DOCUMENT RESUME

ED 462 425 TM 033 689

AUTHOR	Hendrickson, Amy B.; Kolen, Michael J.
TITLE	IRT Equating of the MCAT. MCAT Monograph.
SPONS AGENCY	Association of American Medical Colleges, Washington, DC.
REPORT NO	MCAT-4
PUB DATE	2001-08-00
NOTE	74p.
AVAILABLE FROM	Association of American Medical Colleges, Section for the Medical College Admission Test, 2450 N Street, NW, Washington, DC 20037. Tel: 202-828-0400; Fax: 202-828-1125; Web site: http://www.aamc.org/mcat.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Equated Scores; Higher Education; *Item Response Theory; Models; Test Items; True Scores
IDENTIFIERS	Equipercentile Equating; *Medical College Admission Test

ABSTRACT

This study compared various equating models and procedures for a sample of data from the Medical College Admission Test(MCAT), considering how item response theory (IRT) equating results compare with classical equipercentile results and how the results based on use of various IRT models, observed score versus true score, direct versus linked equating, and various test forms compare. The practical issues and potential benefits of IRT equating are discussed. Data were from 2 forms of the test and 2 administrations, for sample sizes of 8,494, 3,638, 8,147, and 4,478. Choosing between equipercentile and IRT equating impacted MCAT scores. For the Biological Sciences, the effect of using any IRT model would appear to be minimal, but none of the IRT model equivalents for Physical Science and Verbal Reasoning exactly matched those of equipercentile equating. Each IRT model diverged from equipercentile equating in Physical Sciences, with the one-parameter model most discrepant and the three-parameter model least discrepant. IRT model equivalents for the Verbal Reasoning section also diverged from the equipercentile results. Currently, random groups and common-item equating are used for the MCAT, and both of these designs were also considered in this study so that moving to IRT equating need not create design complications, although some statistical assumptions associated with IRT models must be met. Some of the issues in using IRT equating, including choosing a calibration program and IRT model, are discussed. (Contains 12 figures, 37 tables, and 20 references.) (SLD)

ASSOCIATION OF AMERICAN MEDICAL COLLEGES

TM

IRT Equating of the MCAT

MCAT
Monograph 4

August 2001

www.aamc.org/mcat

MCAT AAMC
MEDICAL COLLEGE ADMISSION TEST

IRT Equating of the MCAT©

Amy B. Hendrickson
Michael J. Kolen
University of Iowa

For reprint permission or information on the *MCAT Monograph Series* contact:
John H. Lockwood, Ph.D., Section for the MCAT, AAMC, 2450 N Street NW, Washington DC 20037

Background and Objectives

Large-scale testing organizations are increasingly considering the use of item response theory (IRT) models for test development, scoring, and equating, especially as they ponder the implementation of computerized testing options. As the Association of American Medical Colleges (AAMC) explores the feasibility of a computerized version of the Medical College Admissions Test (MCAT), one of the first issues they are looking at is equating MCAT test forms using IRT methods.

Cook and Eignor (1991) cite several potential theoretical and practical advantages of IRT equating methods compared with conventional methods. They mention the ability of IRT methods to provide conversions that are group invariant and the flexibility that IRT equating affords in choosing previous forms to equate to. However, implementation of IRT equating often requires strong statistical assumptions, such as unidimensionality. Even though such assumptions likely do not hold in practice, IRT equating methods are often found to be robust to violations of these assumptions.

IRT equating has been researched and compared empirically with classical equating methods in many studies. Skaggs and Lissitz (1986) provided a comprehensive review of the literature and issues concerning IRT equating. Among the studies included in the review was Kolen (1981), in which he found that the three-parameter logistic model (Lord, 1980) worked well for equating and performed better than the Rasch in a variety of situations. Kolen (1981) cited a lack of a guessing parameter in the Rasch model as a possible explanation for its lower performance. In other studies reviewed, however, a few pointed to problems with the three-parameter logistic model. Parameter estimation issues were often cited as a possible confounding contributor to these results. Kolen (1981) further found that true score equating produced more stable results than observed score methods, but the procedure used for obtaining true score equivalents below chance level is somewhat arbitrary.

In evaluating the results from different equatings, several criteria can be used. Harris and Crouse (1993) exhaustively reviewed past equating studies in order to summarize these criteria. Although they accumulated an extensive list, they found none to be wholly sufficient or any one to be best. They stated that the situation specific results of equating demand replication and comparison across studies. Choice of criteria is further complicated by the lack of information concerning the true values; results may be compared across equating methods, while neither set of equivalents may resemble the true relationship.

Because of the inexact criteria for assessing equating results, this study will not point to one best model or method. The results, instead, will show how these models compare and contrast. More importantly, they will show for whom and how choice of a model may impact MCAT scores. This study will be useful in informing AAMC as they deliberate on the MCAT, as well as provide a useful guide to issues, concerns, and potential benefits that IRT equating methods may provide.

The purpose of the study is to compare various equating models and procedures for a sample of MCAT data. Specifically, the study will address the following questions:

1.) How do IRT equating results compare with classical equipercentile results?
2.) How do the results, based on use of various IRT models (1-, 2-, and 3-PL), observed score versus true score methods of equating, direct versus linked equating, and various test forms, compare?
3.) What are the practical issues that must be addressed and potential benefits that may be derived if the MCAT moves to IRT equating?

Methodology

The data and procedures used in the study are described. The procedures for conducting equipercentile equating are presented first, followed by those for IRT item parameter estimation, rescaling, and equating. Finally, the criteria used to evaluate and compare the equating results are described.

Data

The data used in the study were drawn from two forms (15 and 23) and two administrations (1994 and 1996) of the MCAT. Data for the Biological Sciences (63 items), Physical Sciences (63 items), and Verbal Reasoning (55 items) test sections were analyzed separately. Item orders a and b--different versions of the same form created by scrambling the order of passage-dependent item sets and non passage-dependent items--were analyzed separately for Form 23. Sample sizes for each form were as follows: 94_15a=8,494; 96_15=3,638; 96_23a=8,147; 96_23b=4,478.

Methods

Classical Equipercentile Equating

Raw-to-scale score conversions for each test section of the 1996 Form 15 data were constructed first by linearly interpolating from the raw-to-scale cut score conversion tables provided by AAMC. These conversions include rounded scale score values, which were derived according to Kolen and Brennan (1995), with procedures similar to those used by ACT in previous equatings of the MCAT. Then, classical equipercentile equatings of Form 23 combined and Forms 23a and 23b separately to the 1996 Form 15 data were conducted with the program RG Equate (Hanson, 1996). Log-linear pre-smoothing degrees (c) were chosen for each distribution based on visual inspection of graphs and evaluations of χ^2 fit and difference statistics. These procedures resulted in both unrounded and rounded scale score equivalents for Forms 23, 23a, and 23b.

IRT Item Parameter Estimation

One-, two-, and three-parameter BILOG analyses were completed for the 96_15, 96_23 complete, and 96_23a and 96_23b individual data. The 'unscrambled' scored data were used such that all item scores were in Order A for all examinees (even for those who took Order B). Missing and 'Not Reached' items, originally coded as '9' in the given data, were recoded as 'Wrong,' indicated by a '0'.

The following BILOG program template was used and modified according to the data file name, sample size, number of items, and number of parameters estimated.

```
MCAT 96_15 BIOLOGICAL SCIENCE
BILOG ONE-PARAMETER MODEL
>GLOBAL  NPArm=1, DFName='C:\96_15-BS.DAT', SAVe;
>SAVe  PARm='c:\96_15-BS1.PAR';
>LENGTH  NITems=63;
>INPUT  NTOtal=63,SAM=3638,NALt=4,NIDC=6;
(6A1,T7,63A1)
>TEST  TNAme=RANDOM;
>CALIB  NQPt=40,FLOat;
```

Notice that the optional FLOAT command was used rather than the BILOG default of no float. Generally, BILOG specifies prior distributions of the parameters and uses these to supplement the information contained in the sample data entered. A default process of the BILOG program uses the means and standard deviations of these pre-specified prior distributions of the item parameter estimates as constants during parameter estimation. This procedure thus influences the values that the parameter estimates may take on, as the estimate of an item parameter is a function of the distance of the parameter value from the mean of its pre-specified distribution, and of the size of the standard deviation of its prior (Baker, 1992). Furthermore, these priors may not be appropriate, even if they prove informative, as they may be very different from the true distributions of the item parameters and thus will pull the item parameter estimates towards inappropriate values.

Use of the FLOAT option in BILOG may help reduce the possible influence of arbitrary, incorrect, or inappropriately specified prior distributions. If the FLOAT option is invoked, the means of the *a* and *c* parameters' priors are estimated by marginal maximum likelihood estimation from the sample's item response data, simultaneously with the item parameters. However, the standard deviations are still fixed. Under the FLOAT option, the means of the distributions are estimated as the average of the n sample item parameters (*a* or *c*) in the test. This estimate is then used as the mean of the prior distribution of each item in the test (Baker, 1992). Estimation of the *b* parameters should not be directly affected by this option. The BILOG manual, as well as other sources, state that the FLOAT option is generally desirable if one has a large sample size and a relatively large number of items, unless one is sure of the appropriate mean values of the prior distributions (Baker, 1992, Mislevy and Bock, 1990).

Forty quadrature points were used instead of the BILOG default of 10. These quadrature points are associated with weights, representing deviates and normalized probability densities of the assumed prior distributions of ability, respectively. The default prior is a Normal distribution. Although 10 quadrature points is the default of BILOG and use of this option may reduce the running time of the program, the theta distribution may be more accurately represented by a larger number of points, say 20 or 40. Using an increased number of quadrature points may help make the smoothed observed score distribution used in the equating procedures even smoother.

Rescaling of IRT Item Parameter Estimates to Common Scale

Of interest in the study was the equating of Form 23 to Form 15 administered in 1996 (96_23), as well as to Form 15 administered in 1994 (94_23). To accomplish this second goal, the Form 23 parameter estimates were put on the same scale as the Form 15 1994 parameter estimates, using a rescaling function to transform the Form 23 parameter estimates to the Form 15 1994 scale. The Stocking-Lord (Stocking and Lord, 1983) parameter rescaling equations were calculated by transforming the 1996 Form 15 parameter estimates to the 1994 Form 15a scale with the program ST (Hanson and Zeng, 1995a). These equations were then used to rescale Form 15 1996 and Form 23 1996.

IRT Equating - True and Observed Score Equating
Form 23 to Form 15 1996

Item parameter estimates and estimated theta distributions from the BILOG analyses were used to compute IRT observed score equating estimates of Form 23 complete and of Forms 23a and 23b, separately, equated to Form 15 administered in 1996, under a randomly equivalent groups design. Only item parameter estimates were used to compute the IRT true score equating relationships. The computer program PIE (Hanson and Zeng, 1995b) was used for both methods of equating. Observed and true score equating are the two methods currently available for conducting IRT equating. In true score equating, number right true scores on Form X are equated to number right true scores on Form Y using the item parameter estimates. Because true scores only range from the sum of the *c* parameters to the number of items on the test, true score equating cannot produce equivalents extending below the limits set by the item pseudochance levels for the three-parameter logistic model.

In observed score equating, however, equivalents are calculated in this region. Observed score equating is conducted by estimating the frequency distributions of number-right observed scores for the two forms and then using ordinary equipercentile equating to approximately equate these estimated observed scores. Although true score equating equivalents are not calculated at the lower end of the score scale, PIE uses an ad hoc procedure to estimate equivalents at these points (Kolen, 1981). Completion of these procedures resulted in both unrounded and rounded scale score equivalents for Forms 23, 23a, and 23b.

Form 23 to Form 15 1994

Form 23 was also equated to the Form 15 1994 data 'through' the Form 15 1996 data. This linked equating presents a second procedure for conducting equating, one more similar to what may be done in practice. For this study, Form 15 administered in 1994 serves as the base form, to which Form 23 should ultimately be equated. This method provides a further consistency check on the results and can provide for more flexible equating designs in the future, as will be described in the discussion. In order to conduct this equating, the Stocking-Lord rescaled 23a, 23b, and 23 complete forms were equated to Form 15a administered in 1994, under a random groups design. These equating analyses were also performed with the program PIE (Hanson and Zeng, 1995b) and resulted in both unrounded and rounded scale score equivalents for Forms 23, 23a, and 23b.

Criteria for Comparing Equating Results

Although no exact equating criteria exist for judging which equating model or procedure is best (Harris and Crouse, 1993), consistency among the results may be observed and assessed. The scale score equivalents were compared between the classical equipercentile and IRT equatings, across the three IRT models, between true and observed score equating methods, between Form 15 1996 and 1994 equatings, and for each form used in the equating. Graphs were created that show the differences between each model's equivalent at each raw score, for both observed and true score equating methods.

Differences in the rounded scale score moments were assessed, as were indices of the weighted differences among the rounded equivalents from each equating. The indices presented are the root mean square (RMS), the mean absolute difference (MAD), and the mean signed difference (MSD). Each index summarizes the discrepancies between equivalents from different equatings at each raw score. The RMS is often used to evaluate statistical error by comparing an estimated value with its true or criterion value. However, as mentioned previously, no true value is known in equating, thus the index will be used merely to compare calculated results across equatings without considering either to be true. Neither the RMS nor the MAD indicates in which direction differences occur, while the MSD does.

Results

The results are organized by test section: Biological Sciences, Physical Sciences, and Verbal Reasoning. Because MCAT scores are reported to examinees as rounded scale scores, these appear to be the results of most interest and are the only ones reported. Comparisons of equivalents across methods and models are reported, rather than results for each method or model separately. The equipercentile results are compared with the IRT equating results first, followed by comparisons between the IRT models, true and observed score equating, and the anchor form that was used for equating. Scale score moments, weighted difference indices, and select individual differences in equivalents across various equating procedures are considered. In the interest of space and because Form 23 is likely to be equated as a whole (instead of by order A and B separately), only results concerning the combined equating of Form 23 are discussed in detail. All tables and figures are included in the appendix.

Biological Sciences

Equipercentile Compared to IRT

Form 23 equated to Form 15 1996 (96_23)

The equipercentile scale score moments for Form 23 equated to Form 15 administered in 1996 are very similar to those from the IRT equating conducted with the BILOG parameter estimates, shown in Table 1. In fact, the moments for the two-parameter model are equivalent to the equipercentile moments, under both observed and true score equating. This equivalence is reflected in the weighted difference indices between the equipercentile results and the two-parameter IRT model results, as the RMS, MAD, and MSD in Table 2 are all .0000. The moments for the one-parameter model are generally higher than the equipercentile and the two-parameter model, while those for the three-parameter model are generally lower. The weighted difference indices show larger differences between the equipercentile results and the three-parameter model than the one-parameter model.

By the general criteria of differences in scale score moments and values of weighted difference indices, the IRT two-parameter model results are most congruent with those for equipercentile equating. The equivalents themselves can be further evaluated to find where and how the equipercentile results differ from the one- and three-parameter models.

Observed Score Equating. From Table 3 it appears that the one-parameter equivalents are one point higher than the other methods' equivalents at raw scores of 13 and 57. The three-parameter equivalent at a raw score of 29 is one point lower than the others. Thus, the one- and three-parameter model rounded equivalents vary from the equipercentile equivalents only at one or two raw score values. Figure 1 shows the 96_23 IRT equivalents as differences from the equipercentile equivalents at each raw score. The bars depicted at the raw scores of 13, 29, and 57 show the discrepancies of the one- and three-parameter models from the equipercentile method.

True Score Equating. Table 4 shows that patterns in true score equating results are similar to those for observed score equating with these data. Differences in the one- and three-parameter models' rounded equivalents occur again at raw scores of 13, 29, and 57. However, at 13, the three-parameter model's equivalent, rather than the one-parameter equivalent as in the observed score equating, is one point higher than the others. Figure 2 depicts these differences.

Form 23 equated to Form 15 1994 (94_23)

In comparing the equipercentile scale score moments with those from IRT equating in which Form 23 was equated back to Form 15 administered in 1994, the findings are similar to those described above. The equipercentile moments are exactly the same as those for the two-parameter model under both observed and true score equating, shown in Table 1. This equivalence is again reflected in the null values of the weighted difference indices in Table 2. The scale score moments and weighted difference indices for the one- and three-parameter models follow the same pattern as above for the 96_23 data.

Observed Score Equating. Table 5 shows that one-parameter equivalents are, again, one point higher than the others at raw scores of 13 and 57. However, the three-parameter model results differ from the other models' more so than in the 96_23 equating. The three-parameter equivalents are one point lower than other models at raw scores of 14, 18, 29, 60, and 62. Figure 3 shows these discrepancies.

True Score Equating. Table 6 shows a pattern of results similar to those for observed score equating. However, the one-parameter model equivalents are one point higher than the others at a raw score of 57 only, while the three-parameter model equivalents are lower than the others at three additional raw scores of 22 and 51. Figure 4 presents these differences.

Table 7 summarizes the number and magnitude of differences between the IRT equating equivalents and the equipercentile equivalents for Biological Sciences.

Comparison of Different IRT Models

Form 23 equated to Form 15 1996 (96_23)

Because the two-parameter model equivalents equal the equipercentile equivalents for these data, differences between equivalents across the IRT models were discussed under the equipercentile heading. However, the general results may be expanded upon. The scale score moments are the same to the tenths place across the models, as shown in Table 1. The weighted difference indices in Table 8a show the one- and three-parameter model equivalents as most different and the one- and two-parameter equivalents as highly similar, for both observed and true score equating. The equivalents differ across IRT models at only 3 raw score points and their pattern is almost identical for observed and true score equating, as shown in Tables 3 and 4.

Form 23 equated to Form 15 1994 (94_23)

Table 1 shows scale score moments for the one- and two-parameter models as the same for the 94_23 equating as for the 96_23 equating. The three-parameter moments changed, however, with seemingly high skewness and kurtosis values under true score equating. Again, the weighted difference indices in Table 8a show one- and three-parameter models as most discrepant and one- and two- parameter models as very similar. The equivalents across models differ at seven (observed) or eight (true) raw score points, shown in Tables 5 and 6.

IRT Observed versus IRT True Score Equating

The Biological Science equivalents are very similar across the two methods of IRT equating. In all reported equatings, the two-parameter model pair is matched. The largest difference between paired observed and true score means is less than .04. Table 8b presents weighted difference indices for observed versus true score equivalents for each IRT model. These values are very low, with the largest differences between the three-parameter model equivalents in the 94_23 equating. Although the scale score moments appear to be very similar across these methods, finding at what values the methods differ is important. For example, Tables 3 and 4 show the observed and true score equivalents for 96_23 differ at a raw score of 13, at the transition from a scale score of 1 to a scale score of 2. For observed score equating, the one-parameter model equivalent is one point higher than others at this point, while for true score equating, the three-parameter equivalent is one point higher. If this transition point is not important, this information is inconsequential; if it is important, then choice of model and equating method are related and the selection becomes more complicated.

Thus, on the level of moments, choice of equating method apparently has only a slight effect, especially for the one- and two-parameter models. However, a closer look at the actual scales reveals possibilities for greater effects on some examinees' scores.

IRT Direct versus IRT Linked Equating

Results of the 96_23 and 94_23 equatings differed only slightly for the Biological Sciences section. The scale score moments in Table 1 show consistent results for one- and two-parameter models across the two equating procedures. Table 8c presents weighted difference indices between 96_23 and 94_23 equivalents and reflects the similarity in results. The largest differences are between the three-parameter model equivalents under true score equating.

Physical Sciences

Equipercentile Compared to IRT

Form 23 equated to Form 15 1996 (96_23)

The equipercentile scale score moments for Form 23 equated to Form 15 administered in 1996 are similar to those from IRT equating conducted with BILOG parameter estimates, as shown in Table 9. None of the IRT models exactly replicate the equipercentile results, but the two-parameter model has the closest mean value. The weighted difference indices in Table 10 point to the largest difference as that between the equipercentile results and the one-parameter IRT model results.

The equivalents can be further evaluated to find where and how equipercentile results differ from the IRT models' equivalents.

Observed Score Equating. From Table 11 it appears that the one-parameter equivalents are one point lower than the other methods' equivalents at raw scores of 11, 19, 24, 56, 58, and 61. The one- and two-parameter models' equivalents are one point lower than the three-parameter and equipercentile equivalents at a score of 14, and one point higher at 38 and 43. Finally, all three IRT models' equivalents are lower than those of equipercentile equating at raw scores of 10, 23, 28, and 53. Thus, the one-parameter model rounded equivalents vary the most from the other equivalents, including those of equipercentile equating, but none of the models yield concordant results. Figure 5 shows the 96_23 IRT equivalents as differences from the equipercentile scale scores at each raw score. The discrepancies between IRT models and the equipercentile method are shown by bars depicted at raw scores of 10, 11, 14, 19, 23, 24, 28, 38, 43, 53, 56, 58, and 61.

True Score Equating. The patterns in true score equating results are similar to those for observed score equating, as shown in Table 12. The only differences from observed score equating results are changes at raw scores of 11, 15, and 61. At 11, both the one- and two-parameter models' rounded equivalents are now 1 point lower than for the other models. At 15, the equivalent of the one-parameter model is 1 point lower than for the other models, but the equivalent at a raw score of 61 for this model is now equal to the other models' equivalents. Figure 6 presents these results.

Form 23 equated to Form 15 1994 (94_23)

In comparing the equipercentile scale score moments to those from IRT equating in which Form 23 was equated back to Form 15 administered in 1994, the findings are similar to those described above, as shown in Table 9. The weighted difference indices in Table 10 vary in pointing to the model closest to the equipercentile equivalents. By RMS and MAD statistics, the three-parameter model is more similar, but by MSD, the results vary.

Observed Score Equating. Table 13 shows the pattern of equivalents discussed for the 96_23 equating holds true for the 94_23 equating. Differences in equivalents occur at the same raw score values and in the same directions, for the most part. One distinction from the 96_23 results is a one point drop in three-parameter equivalents at raw scores of 11, 14, 56, and 58. These changes bring greater consistency between one- and three-parameter models and less similarity between the equivalents of the three-parameter and equipercentile equatings. Figure 7 depicts these results.

True Score Equating. A similar pattern of results as for observed score equating appears, again with changes to three-parameter equivalents compared with those in 96_23 equating. Table 14 and Figure 8 present these results.

Table 15 summarizes the number and magnitude of differences between IRT equating equivalents and equipercentile equivalents for Physical Sciences.

IRT Results Compared for Different IRT Models

Form 23 equated to Form 15 1996 (96_23)

The first two scale score moments are the same to the tenths place across the models, while the third and fourth moments tend to vary more across models, as depicted in Table 9. The RMS and MAD indices in Table 16a show one- and three-parameter model equivalents as most different and one- and two-parameter equivalents as similar, for both observed and true score equating. The equivalents differ across IRT models at 13 points. This pattern is almost identical for both observed and true score equating, as shown in Tables 11 and 12.

Form 23 equated to Form 15 1994 (94_23)

The scale score moments for one- and two-parameter models under true score equating are the same for 94_23 equating as they were for 96_23 equating. Weighted difference indices in Table 16a show one- and three-parameter models and two- and three-parameter models as most discrepant. The equivalents across models differ at seven raw score points, generally where the one- or three-parameter model equivalents are one point lower than the others, as shown in Tables 13 and 14.

IRT Observed versus IRT True Score Equating

Table 16b presents weighted difference indices for observed versus true score equivalents for each IRT model and shows that the Physical Science equivalents are less consistent across the two methods of IRT equating than were the Biological Science equivalents. In fact, for 94_23, none of the three pairs of results between observed and true score equating match. In the other two equating sets, only one of three model pairs is matched. However, the matched pair is the three-parameter model equivalents for 96_23 equating, which did not match for any equating set with the Biological Sciences data. More matched pairs appear between the forms equated to (i.e., between 96_23 and 94_23) than between equating methods.

IRT Direct versus IRT Linked Equating

The results of the 96_23 and 94_23 equatings differ only slightly for the Physical Sciences section. The scale score moments in Table 9 show consistent results for one- and two-parameter models across the two equating procedures, under true score equating. Table 16c presents weighted difference indices between 96_23 and 94_23 equivalents and reflects the similarity in the results. The largest differences are between the three-parameter model equivalents under observed score equating.

Verbal Reasoning

Equipercentile Compared to IRT

Form 23 equated to Form 15 1996 (96_23)

Table 17 shows that equipercentile scale score moments for Form 23 equated to Form 15 administered in 1996 are slightly lower than those from IRT equating conducted with BILOG parameter estimates. None of the IRT models exactly replicate the equipercentile results, but moments of the three-parameter model under observed score equating are similar. Two of the weighted difference indices in Table 18 point to the largest difference as that between equipercentile results and the one-parameter IRT model results. The MSD statistic implicates the two- and three-parameter models as most discrepant from equipercentile under observed and true score equating, respectively.

The equivalents can be further evaluated to find where and how equipercentile results differ from the IRT models' equivalents.

Observed Score Equating. From Table 19 it appears that the one-parameter equivalents are one point lower than the other methods' equivalents at a raw score of 18, and one- and two-parameter models' equivalents are one point lower than the three-parameter and equipercentile equivalents at scores of 12, 13, and 21, and one point higher at a score of 32. The two-parameter equivalent at a raw score of 52 is one point higher than other models. Finally, all three IRT models' equivalents are lower than those of the equipercentile equating at raw scores of 17, 48, and 50, and one point higher at 35. Thus, the one-and two-parameter models' rounded equivalents vary equally from three-parameter and equipercentile equivalents. Figure 9 shows 96_23 IRT equivalents as differences from equipercentile equivalents at each raw score. The discrepancies of the IRT models from the equipercentile method are shown by the bars depicted at the raw scores of 12, 13, 17, 18, 21, 32, 35, 48, 50, and 52.

True Score Equating. Many of the model discrepancies found in observed score equating results also appear in true score equating results, with several additions shown in Table 20. Differences in true score equating models are also found at raw scores of 11, 14, 25, and 28, due mostly to changes in the one- and three-parameter models' equivalents from observed score to true score equating. From this analysis, the one-parameter model differs from the others at 4 points, the one- and two-parameter models at 6 points, the three-parameter at one point, the one- and three-parameter models at one point, and finally, all IRT models differ from the equipercentile equivalents at one point. Figure 10 presents these results.

Form 23 equated to Form 15 1994 (94_23)

In comparing the equipercentile scale score moments with those from IRT equating in which Form 23 was equated back to Form 15 administered in 1994, the findings are similar to those described above. In fact, Table 17 shows that the one-parameter moments for 94_23 equating match those of 96_23 equating for both observed and true score methods. The weighted difference indices in Table 18 vary in pointing to the model closest to the equipercentile equivalents. By the RMS and MAD statistics, the three-parameter model is more similar; by MSD, however, the one-parameter model's equivalents are most like those resulting from equipercentile equating.

Observed Score Equating. Table 21 shows that the pattern of equivalents discussed for 96_23 observed score equating holds true for 94_23 observed score equating. Differences in equivalents occur at the same raw score values and in the same directions, for the most part. One distinction from 96_23 results is a one point drop in three-parameter equivalents at raw scores of 12, 21, and 32. These changes bring greater consistency between the results of the IRT models. Figure 11 depicts these results.

True Score Equating. A similar pattern of results appears as for 94_23 observed score equating, with changes again to three-parameter equivalents compared with those in the 96_23 equating. Table 22 and Figure 12 present these results.

Table 23 summarizes the number and magnitude of differences between IRT equating equivalents and equipercentile equivalents for Verbal Reasoning.

IRT Results Compared for Different IRT Models

Form 23 equated to Form 15 1996 (96_23)

Table 17 shows that the first two scale score moments are the same to the tenths place across models, while the third and fourth moments tend to vary more. The RMS and MAD indices in Table 24a show the one- and three-parameter model equivalents as most different and the one- and two-parameter or two- and three-parameter equivalents as similar. The equivalents differ across IRT models at 6 raw score points under observed score equating, as shown in Table 19. Table 20 shows a similar pattern for true score equating, with a few additional points where one or more of the equating models diverge from the original scale score. For example, at a raw score of 14, the one-parameter model equivalent is one point lower than the equipercentile equated scale score.

Form 23 equated to Form 15 1994 (94_23)

Scale score moments for the one-parameter model in Table 17 are the same for the 94_23 equating as they were for the 96_23 equating, for both observed and true score equating. The weighted difference indices in Table 24a, again, generally show the one- and three-parameter models as most discrepant. The equivalents differ across models at only two raw score values for observed score equating, but at six points for true score equating, generally where three-parameter model equivalents are one point higher than the others. These results are seen in Tables 21 and 22.

IRT Observed versus IRT True Score Equating

Table 24b presents weighted difference indices for observed versus true score equivalents for each IRT model and shows that none of the equatings presented led to consistent results across the two methods of IRT equating, also shown in Table 17. The values in Table 24b are moderately high, with the largest differences between the three-parameter model equivalents in 96_23 equating. Again, more matched pairs occurred between the forms equated to (i.e., between 96_23 and 94_23) than between equating methods. For example, the one-parameter true score equating results for 94_23 more closely resemble the results of the true score equating of 96_23 than they do the observed score equating for 94_23.

IRT Direct versus IRT Linked Equating

Results of the 96_23 and 94_23 equatings differed some for the Verbal Reasoning section. Scale score moments in Table 17 show consistent results for the one-parameter model across the two equating procedures. Table 24c presents weighted difference indices between 96_23 and 94_23 equivalents and reflects the similarity in the results. The largest differences are between the three-parameter model equivalents under true score equating.

Discussion

As mentioned previously, no exact criteria exist for evaluating and comparing equating results. Harris and Crouse (1993) found that different criteria may lead to the choice of different methods, thus no one criteria may be stated as best, and more than one equating method may be adequate.

Classical Equipercentile Equating versus IRT Equating

In the current study, choosing between equipercentile and IRT equating could impact MCAT scores. For the Biological Sciences, the effect of using any IRT model would appear to be minimal; in fact, choosing the two-parameter logistic model instead of equipercentile would not change the equating results for this sample. However, none of the IRT models' equivalents for the Physical Science and Verbal Reasoning sections exactly match those of the equipercentile. Each of the IRT models diverge from equipercentile in Physical Sciences, but the one-parameter model is indicated as most discrepant and the three-parameter as least. While one-parameter equivalents are generally lower than the equipercentile at an average of 13 raw score points, the two- and three-parameter equivalents differ only by an average of 8 and 6 points, respectively. These points fall across the score distribution and could impact scores of students at all abilities.

IRT model equivalents for the Verbal Reasoning section also diverge from the equipercentile results. Again, the one-parameter model is indicated as most discrepant and the three-parameter as most congruent. The one- and two-parameter model equivalents are generally lower than the equipercentile at an average of 11 and 9 raw score points, respectively, while the three-parameter equivalents differ only by an average of 5 points. These differences fall across the score distribution from scores of 12 to 52. Again, whether we ideally want to replicate equipercentile results or not is unknown.

Currently, two designs are used to equate the MCAT, random groups and common-item. Both designs were also used in this study of MCAT IRT equating; thus, moving to IRT equating need not create new design complications. However, additional statistical assumptions associated with item response theory models must be met. Some of these assumptions may be addressed in other MCAT Monographs. Further, if either of the models derived from an IRT calibrated item pool are to be used, to be described later, representative common-item sets need to be constructed, and context effects and multidimensionality need to be assessed. Although a common-item design has been previously used, all items were the same between the forms, thus eliminating the concern for representativeness and context effects appear to not have been a problem. Table 8.1 in Kolen and Brennan (1995) compares these designs in terms of complications in test administration and development as well as the statistical assumptions required.

The classical equipercentile, Rasch, and three-parameter logistic methods of equating are appropriate for many of the same situations under both random groups and common-item designs (see Table 8.5 in Kolen and Brennan, 1995). However, use of the Rasch or three-parameter methods also require that assumptions of the IRT model are reasonably met and, for the three-parameter model, that the increased computational demands are feasible. The first of these issues will be addressed in a future Monograph; the second is easily addressed with the use of current computer programs.

Issues in Implementing IRT Equating

If IRT equating were to be implemented, several decisions would have to be made; results from this study may help inform some of them. First, because equating is so dependent on estimation of item parameters, choice of a calibration program can affect equating results. These programs follow different procedures and provide various options. A decision to use BILOG, MULTILOG, or even LOGIST or PARSCALE, must be made, along with choosing options within each program. Equating can be conducted with parameters from each of these programs and with various options before a decision is made.

Second, choices may be made concerning the transformation function used for equatings that involve common-item design. The Stocking-Lord transformation was used in this study, but other procedures exist, namely the Mean/Mean, Mean/Sigma, and Haebara methods. Both the Stocking-Lord and Haebara functions are characteristic curve methods in which all item parameters are considered simultaneously in calculating rescaling equations. These methods have been found to be at least--and often more--accurate than the other two methods, in which only the means and standard deviations of item parameters are considered in calculating transformation functions (Baker and Al-Karni, 1991; Hung, Wu, and Chen, 1991; Way and Tang, 1991). However, it is recommended that equating be conducted with parameters derived from each of the transformation methods, and that the results, including differences in scale scores, be compared before choosing which to report.

A choice of IRT model to use in equating is also important. As with choosing between equipercentile and IRT equating, the effect of various IRT models for this sample was found to vary by test section. For Biological Sciences, the decision had only slight ramifications for equating results and thus for test scores. However, for Physical Sciences and Verbal Reasoning, the IRT equivalents varied from each other across the score distribution. The one- and three-parameter models were generally indicated as most discrepant from each other in both the Physical Sciences and Verbal Reasoning sections, and the one- and two-parameter models were most similar. Which model is best is unknown. Closer inspection of where and how the equivalents differed at particular points on the distribution may help decide which is most appropriate for the MCAT.

A choice between true and observed score equating methods also needs to be made and both have advantages and disadvantages. True score equating procedures are easier computationally and do not depend on the ability distribution of examinees. However, this method requires the use of true scores that are unobtainable in practice; equivalents cannot be calculated at raw scores that are less than or equal to the sum of the *c* parameters or that correspond to the top number correct score. Observed score equating, on the other hand, relies only on the availability of examinees' observed scores; its increased computational complexities are feasible if the posterior theta distribution from the IRT calibration program is used. The largest differences between these methods will generally be at the low and high end of scores because of the lack of equivalents from the true score method at these points. In the current study, the program used for true score equating, PIE, incorporated an ad hoc procedure for calculating equivalents in these regions (Kolen, 1981) and differences between methods were not generally found in these areas.

The results of this study showed consistency between observed score and true score methods of equating, as shown in Tables 8b, 16b, and 24b. However, choice of an equating method may depend on the IRT model chosen. For example, in Biological Sciences, under 96_23 equating, the two-parameter model's results are the same across observed and true score equating. However, under observed score equating, the one-parameter equivalent at a raw score of 13 is one point higher than the others, while under true score equating, the three-parameter equivalent is one-point higher than the others. Thus, choice of observed versus true score equating may be related to choice of an IRT model and may impact examinees' scores. Using both IRT true score and observed score methods is recommended, as is comparing equating relationships and scale score differences before choosing which to report.

A final issue in implementing IRT equating involves the use of the common-item design. In using the common-item design with IRT equating, as was done to equate Form 23 to Form 15 administered in 1994 and which would be used for equating scrambled forms, the invariance of the item parameter estimates between the forms/administrations must be assessed. This can be done by plotting the estimates for one form against the other and looking for outliers. If any items' parameter estimates appear to vary between the two forms, equating should be conducted with and without these items, and the results compared. This design can still be used and equating can continue, but these items may not be included as common items if they irregularly affect the equating results. In this study, analyses of parameter invariance between Form 15 administered in 1994 and Form 15 administered in 1996 showed highly consistent estimates; thus, all items were included as common.

Potential Future Uses of IRT Equating

Once IRT equating has been implemented and used for several years, two potential test development and equating designs may be derived from its use.

IRT Calibrated Item Pool

These pools are groups of items that contain item parameter estimates all placed on the same ability scale. Using the pools may provide greater flexibility in developing test forms. To create a pool of IRT calibrated items, Form Y is administered, and then scale scores and IRT parameters are calculated. Next, Form X1, containing some old items from Form Y and some new items, is administered. Common-item equating is conducted to place Form X1 scores on the Form Y score scale. Random groups equating between Form Y and Form X1 without common items may also be used, which is how MCAT is currently equated. Now an item bank exists consisting of Form Y only, Form X1 only, and Form Y and Form X1 common items on the same ability scale. From here, a new form, X2, can be constructed with some old items from the pool and some new items. Common-item equating to the pool is conducted to place the Form X2 scores onto the same scale. (Form X2 scale scores are obtained from the Form Y raw-to-scale score conversion.) In this way, items on the same ability scale will be continually added to the pool. Future tests may include any representative sample of items from the pool, rather than from just one past test form.

In this study, the equating of Form 23 to Form 15 1994 followed the essential steps of creating an item pool. The items from these forms have now been transformed to the same ability scale and may be considered an IRT calibrated item pool that could be used in future test development and equating. The observed invariance of item parameters across Form 15 1994 and Form 15 1996, and the highly consistent results of the 96_23 and 94_23 equatings, as shown in Tables 8d, 16d, and 24d, indicate the linked equating method as a viable option for the MCAT.

There are several issues involved in using an IRT calibrated item pool. First, items may need to be removed from the pool as their content becomes dated or for security reasons. Second, with repeated administration of items, parameter estimates are likely to change. For example, although the parameter estimates of the two administrations of Form 15 are highly correlated, the values did change from 1994 to 1996. A decision must be made about how to account for these changes. McKinley (1988) evaluated six methods for combining item parameter estimates over administrations. These methods were the unweighted average, sample-size weighted average, standard error weighted average, covariance matrix weighted average, sample size weighted geometric average, and partial weighted average. He recommended using the covariance matrix weighted average and partial weighted average procedures.

Context effects are another concern: whereby the order in which items are administered influences their estimated parameter values, and problems result from using items written under different test specifications. And as with any test to be equated under a common-item design, the number and characteristics of the items chosen from the pool must adequately represent the total test.

Finally, classical equipercentile equating typically cannot be conducted for test forms created from an IRT calibrated item pool. Thus, before considering the use of a calibrated item pool, it is recommended that the simple IRT common-item design be used for several administrations in order to test the IRT assumptions and compare the results with traditional methods.

Item Preequating

Another similar derivation of IRT equating is item preequating, which allows for construction of raw-to-scale score conversion tables before forms are administered. Through this design, examinees' tests contain items that have been previously administered and calibrated. Examinees' scores can be converted to scale scores and reported back without waiting for equating to be conducted. Item preequating may also be helpful in dealing with test disclosure legislation. As will be described, under a preequating design the new items on each test form do not contribute to examinees' scores, thus they do not need to be disclosed until they are included as operational items on future forms.

Development of an item preequating design is similar to that of an IRT calibrated item pool. First, Form Y is created, which contains both operational and non-operational items that do or do not contribute to an examinee's score. A scale score conversion table is created for the operational items, the entire test is administered, and item parameters are calculated for all items. At this point, an IRT calibrated item pool has been developed, consisting of both operational and non-operational items. Then a new form, X1, is created, containing operational items from the pool and new non-operational items. In this way, item parameter estimates already exist for the operational items of Form X1 (i.e., they are 'preequated') and a conversion table can be created before it is administered. After administration, all Form X1 items are calibrated and the operational items are used as a common-item set for transforming new non-operational items to the Form Y ability scale. These new items can then be added to the pool.

As with the use of an IRT calibrated pool, several issues must be considered. First, if a miskey is discovered during administration, the preequated conversion table must be adjusted. Decisions about removing items from the pool must be made, as well as accounting for changes in item parameter estimates over administrations (McKinley, 1988). Finally, and most importantly, the possibility of both context effects and multidimensionality complicate the use of preequating. In studies of the ACT and SAT, problems arose when operational items were presented in a different context or position than they were as non-operational items. Furthermore, if non-operational items are not representative of the total test content, estimation of parameters for all items is more difficult. In this situation, the possibility of multidimensionality exists if non-operational items are calibrated with operational items. These issues must be addressed if item preequating is to be implemented.

Comparison of these Designs

The IRT calibrated pool and item preequating designs differ on whether new items contribute to an examinee's score and if the conversion table can be constructed before the new form is administered. Figure 6.9 in Kolen and Brennan (1995) depicts the similarities and differences in these designs.

Computerized Tests

In a review of studies comparing computerized versus paper-and-pencil tests, Mazzeo and Harvey (1988) recommended that separate equatings and normings be conducted for each. They pointed to differences in scores across these modes as support for their recommendation. Of particular concern was the equivalence of tests that included reading passages and graphics, or that were speeded. The equivalence of scores on computer-adaptive and paper-and-pencil tests is even less substantiated than for simple computerized tests. The two equating designs described above, IRT calibrated item pools and item preequating, form the basis for computerized adaptive tests (CATs). As is the case when these designs are used for paper-and-pencil tests, newly administered items may or may not contribute to examinees' scores. Wainer and Mislevy (1990) describe this second situation, called on-line calibration.

The concerns described for IRT calibrated pools and item preequating are even more complex for CATs. For example, context effects are of great concern, for each item in a CAT may be presented in a different position and/or context for each examinee or administration. Also, currently, many of the items on the MCAT are contained within passage subsets and use of a CAT that contains stimuli-based items would be complicated. Context effects can occur within and between these passage sets, and how these alterations affect item parameter estimates still requires substantial research. The forthcoming Monograph regarding local item dependence may shed further light on this issue.

In a synthesis of research, Kolen (1999-2000) cited several concerns regarding the equivalence of computerized and paper-and-pencil assessments. First, because of the adaptive nature of CATs, the content and difficulty of test questions administered to examinees may vary. Tests of different content may be assessing different constructs, thus content-balancing is necessary. Several procedures have been developed and do appear to constrain this issue. Second, differences in statistical specifications between paper-and-pencil and CATs may affect examinees, both individually and as groups. For example, conditional standard errors of measurement may vary, affecting the equivalence of scores across these administration modes. Kolen (1999-2000) cited other potential threats to paper-and-pencil and CAT score equivalence due to differences during and after test administration, such as testing conditions and scoring. He also described issues to consider when using alternate pools of computerized adaptive tests. Each of these issues must be addressed if the MCAT is to move to computerized adaptive testing.

Conclusions

"Does it matter if IRT equating is used for the MCAT?" The answer is clearly yes. For this sample, the effect of IRT equating on MCAT scores compared with the use of equipercentile equating depended on the test section, the IRT model chosen, and the equating method used. Overall, equivalents between the equating procedures only differed by one point, higher or lower, at various points across the score distribution. Decisions can be made regarding: 1) at what scores these differences are most important, and 2) a model and method that leads to the most appropriate results at these scores. A study that replicates these results, as well as studies with other item calibration programs and scale transformation procedures, should be undertaken.

References

Baker, F.B. (1992). *Item response theory parameter estimation techniques*. New York: Marcel Dekker.

Baker, F.B., and Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147-162.

Childs, R.A., and Oppler, S.H. (1999). *Practical implications of subtest dimensionality for item response calibration of the Medical College Admissions Test.* Washington, DC: American Institutes for Research.

Cook, L., and Eignor, D. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10(3)*, 37-45.

Hanson, B.A. (1996). *RG Equate: A program for smoothed equipercentile equating using the random groups design, Version 1.1.1.* Iowa City, IA: ACT.

Hanson, B.A., and Zeng, L. (1995a). *ST: A computer program for IRT scale transformation, Version 1.0.* Iowa City, IA: ACT.

Hanson, B.A., and Zeng, L. (1995b). *PIE: A computer program for IRT equating, Version 1.0.* Iowa City, IA: ACT.

Harris, D.J., and Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195-240.

Hung, P., Wu, Y., and Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking.* Paper presented at the International Academic Symposium on Psychological Measurement.

Kolen, M.J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6(2)*, 73-96.

Kolen, M.J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 22,* 197-206.

Kolen, M.J., and Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mazzeo, J., and Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Report 88-8). New York: College Entrance Examination Board.

McKinley, R.L. (1988). A comparison of six methods for combining multiple IRT item parameter estimates, *Journal of Educational Measurement, 25(3)*, 233-246.

Mislevy, R.J., and Bock, R.D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Skaggs, G., and Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495-529.

Stocking, M.L., and Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Wainer, H., and Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (ed.), *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Erlbaum.

Way, W.D., and Tang, K.L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Table 1. Scale Score Moments of Biological Sciences Equivalents

			Mean	SD	Skewness	Kurtosis
Equipercent.						
			8.1394	**2.2919**	**-0.2780**	**2.9676** ^{1}
IRT	96_23	Observed				
		1-PL	**8.1497**	**2.3087**	**-0.2340**	**2.9782** ^{2}
		2-PL	**8.1394**	**2.2919**	**-0.2780**	**2.9676** ^{1}
		3-PL	8.1180	2.3163	-0.2780	2.8903
		True				
		1-PL	**8.1491**	**2.3103**	**-0.2388**	**2.9928** ^{3}
		2-PL	**8.1394**	**2.2919**	**-0.2780**	**2.9676** ^{1}
		3-PL	8.1186	2.3147	-0.2734	2.8760
	94_23	Observed				
		1-PL	**8.1497**	**2.3087**	**-0.2340**	**2.9782** ^{2}
		2-PL	**8.1394**	**2.2919**	**-0.2780**	**2.9676** ^{1}
		3-PL	8.1086	2.3189	-0.3404	2.9117
		True				
		1-PL	**8.1491**	**2.3103**	**-0.2388**	**2.9928** ^{3}
		2-PL	**8.1394**	**2.2919**	**-0.2780**	**2.9676** ^{1}
		3-PL	8.0720	2.3164	-0.3873	3.0111

Note: Highlighted cells with superscript numbers indicate matched scale score moments within or across equatings. For example, the RG Equate moments and the 2-PL IRT moments are equivalent.

Table 2. Weight. Diff. Indices between Equipercentile and IRT Biological Sciences

		RMS	MAD	MSD	
96_23	Observed				
	EQU-1PL	0.1015	0.0103	0.0103	
	EQU-2PL	*0.0000*	*0.0000*	*0.0000*	1
	EQU-3PL	**0.1462**	**0.0214**	**0.0214**	
	True				
	EQU-1PL	0.0987	0.0097	-0.0097	
	EQU-2PL	*0.0000*	*0.0000*	*0.0000*	1
	EQU-3PL	**0.1481**	**0.0219**	**0.0208**	
94_23	Observed				
	EQU-1PL	**0.2067**	**0.0427**	0.0238	
	EQU-2PL	*0.0000*	*0.0000*	*0.0000*	1
	EQU-3PL	0.1824	0.0333	**0.0333**	
	True				
	EQU-1PL	0.2065	0.0423	0.0235	
	EQU-2PL	*0.0000*	*0.0000*	0.0000	1
	EQU-3PL	**0.2182**	**0.0439**	**0.0433**	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 7. Number of -1 and +1 point differences between IRT rounded scale score equivalents and equipercentile equivalents - Biological Sciences

Section	Form	Method		-1	1
BS	96_23	Observed			
			1-PL	0	2
			2-PL	0	0
			3-PL	1	0
		True			
			1-PL	0	1
			2-PL	0	0
			3-PL	1	1
	94_23	Observed			
			1-PL	0	2
			2-PL	0	0
			3-PL	5	0
		True			
			1-PL	0	1
			2-PL	0	0
			3-PL	7	0

Table 8. Weighted Difference Indices for Biological Sciences Equivalents

				RMS	MAD	MSD	
8a.	IRT	96_23	Observed				
			1PL-2PL	*0.1015*	*0.0103*	*0.0103*	
			1PL-3PL	**0.1780** [1]	**0.0317** [2]	**0.0317**	
			2PL-3PL	0.1462	0.0214	0.0214	
			True				
			1PL-2PL	*0.0987*	*0.0097*	*0.0097*	
			1PL-3PL	**0.1780** [1]	**0.0317** [2]	**0.0306**	
			2PL-3PL	0.1481	0.0219	0.0208	
		94_23	Observed				
			1PL-2PL	*0.1015*	*0.0103*	*0.0103*	
			1PL-3PL	**0.2028**	**0.0411**	**0.0411**	
			2PL-3PL	0.1755	0.0308	0.0308	
			True				
			1PL-2PL	*0.0987*	*0.0097*	*0.0097*	
			1PL-3PL	**0.2778**	**0.0771**	**0.0771**	
			2PL-3PL	0.2596	0.0674	0.0674	
8b.	Observe	96_23					
	- True		**1PL-1PL**	**0.0235**	**0.0006**	**0.0006**	[3]
			2PL-2PL	0.0000	0.0000	0.0000	[4]
			3PL-3PL	**0.0235**	**0.0006**	**-0.0006**	[3]
		94_23					
			1PL-1PL	0.0235	0.0006	0.0006	[3]
			2PL-2PL	*0.0000*	*0.0000*	*0.0000*	[4]
			3PL-3PL	**0.1913**	**0.0366**	**0.0366**	
8c.	96-94	Observed					
			1PL-1PL	*0.0000*	*0.0000*	*0.0000*	[4]
			2PL-2PL	*0.0000*	*0.0000*	*0.0000*	[4]
			3PL-3PL	**0.0971**	**0.0094**	**0.0094**	
		True					
			1PL-1PL	*0.0000*	*0.0000*	*0.0000*	[4]
			2PL-2PL	*0.0000*	*0.0000*	*0.0000*	[4]
			3PL-3PL	**0.2158**	**0.0466**	**0.0466**	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 9. Scale Score Moments of Physical Sciences Equivalents

				Mean	SD	Skewness	Kurtosis
Equipercent.							
				7.9176	2.7711	0.3834	2.8610
IRT	96_23	Observed					
			1-PL	7.8913	2.2835	0.0904	2.6511
			2-PL	7.9287	2.2879	0.2212	2.7890
			3-PL	7.8623	2.2812	0.3148 ^{1}	2.8203 ^{2}
		True					
			1-PL	7.8909	2.2894	0.0908	2.6950 ^{3}
			2-PL	7.9230	2.2989	0.1973	2.8114 ^{4}
			3-PL	7.8623	2.2812	0.3148 ^{1}	2.8203 ^{2}
	94_23	Observed					
			1-PL	7.8923	2.2862	0.0993	2.6786
			2-PL	7.9231	2.2987	0.1980	2.8095
			3-PL	7.8448	2.2494	0.2349	2.7079
		True					
			1-PL	7.8909	2.2894	0.0908	2.6950 ^{3}
			2-PL	7.9230	2.2989	0.1973	2.8114 ^{4}
			3-PL	7.8486	2.2856	0.2417	2.7998

Note: Highlighted cells with superscript numbers indicate matched scale score moments within or across equatings. For example, 96_23 1-PL moments match for observed and true score equating.

Table 10. Weight. Diff. Indices between Equipercentile and IRT Physical Sciences

		RMS	MAD	MSD	
96_23	Observed				
	EQU-1PL	**0.4024**	**0.1619**	0.0263	
	EQU-2PL	0.3529	0.1245	*-0.0111*	
	EQU-3PL	*0.2353*	*0.0554*	**0.0554**	1
	True				
	EQU-1PL	**0.4029**	**0.1623**	**0.1623**	
	EQU-2PL	0.3609	0.1302	*-0.0054*	
	EQU-3PL	*0.2353*	*0.0554*	0.0554	1
94_23	Observed				
	EQU-1PL	**0.3544**	**0.1256**	*-0.0090*	
	EQU-2PL	0.2006	0.0403	**0.0403**	2
	EQU-3PL	*0.1677*	*0.0281*	0.0257	
	True				
	EQU-1PL	**0.3530**	**0.1246**	**0.0527**	
	EQU-2PL	0.2006	0.0403	0.0403	2
	EQU-3PL	*0.1511*	*0.0228*	*0.0204*	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 15. Number of -1 and +1 point differences between IRT rounded scale score equivalents and equipercentile equivalents - Physical Sciences

Section	Form	Method		-1	1
PS	96_23	Observed			
			1-PL	11	2
			2-PL	5	2
			3-PL	4	0
		True			
			1-PL	11	2
			2-PL	7	2
			3-PL	4	0
	94_23	Observed			
			1-PL	10	2
			2-PL	6	2
			3-PL	8	0
		True			
			1-PL	11	2
			2-PL	7	2
			3-PL	9	0

Table 16. Weighted Difference Indices for Physical Sciences Equivalents

				RMS	MAD	MSD	
16a.	IRT	96_23	Observed				
			1PL-2PL	*0.1934*	*0.0374*	-0.0374	
			1PL-3PL	**0.3264**	**0.1065**	*0.0291*	
			2PL-3PL	0.2630	0.0691	**0.0665**	
			True				
			1PL-2PL	*0.1791*	*0.0321*	-0.0321	1
			1PL-3PL	**0.3270**	**0.1069**	*0.0287*	
			2PL-3PL	0.2736	0.0749	**0.0608**	
		94_23	Observed				
			1PL-2PL	*0.1755*	*0.0308*	*-0.0308*	
			1PL-3PL	0.2968	0.0881	0.0475	
			2PL-3PL	**0.2993**	**0.0896**	**0.0783**	
			True				
			1PL-2PL	*0.1791*	*0.0321*	*-0.0321*	1
			1PL-3PL	**0.3055**	**0.0933**	0.0423	
			2PL-3PL	0.2727	0.0744	**0.0744**	
16b.	Observe - True	96_23					
			1PL-1PL	0.0479	0.0023	0.0004	
			2PL-2PL	**0.0755**	**0.0057**	**0.0057**	
			3PL-3PL	*0.0000*	*0.0000*	*0.0000*	2
		94_23					
			1PL-1PL	0.0367	0.0013	0.0013	
			2PL-2PL	*0.0089*	*0.0001*	*0.0001*	
			3PL-3PL	**0.1335**	**0.0178**	**-0.0039**	
16c.	96-	Observed					
			1PL-1PL	0.0308	0.0010	-0.0010	
			2PL-2PL	*0.0750*	*0.0056*	*0.0056*	
			3PL-3PL	**0.1323**	**0.0175**	**0.0175**	
		True					
			1PL-1PL	*0.0000*	*0.0000*	*0.0000*	2
			2PL-2PL	*0.0000*	*0.0000*	*0.0000*	2
			3PL-3PL	**0.1167**	**0.0136**	**0.0136**	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 17. Scale Score Moments of Verbal Reasoning Equivalents

				Mean	SD	Skewness	Kurtosis
Equipercent.				7.8109	2.3321	-0.1087	2.5998
IRT	96_23	Observed					
			1-PL	7.8539	2.2973	-0.3734	2.8792[1]
			2-PL	7.8615	2.2898	-0.3307	2.8548
			3-PL	7.8327	2.2839	-0.2670	2.6421
		True					
			1-PL	7.8070	2.3554	-0.3868	2.7776[2]
			2-PL	7.8318	2.3440	-0.2910	2.7620
			3-PL	7.7760	2.3326	-0.1348	2.4724
	94_23	Observed					
			1-PL	7.8539	2.2973	-0.3734	2.8792[1]
			2-PL	7.8594	2.2845	-0.3474	2.8232
			3-PL	7.8607	2.2810	-0.3377	2.7930
		True					
			1-PL	7.8070	2.3554	-0.3868	2.7776[2]
			2-PL	7.7712	2.3706	-0.3309	2.6322
			3-PL	7.8228	2.3031	-0.2967	2.6813

Note: Highlighted cells with superscript numbers indicate matched scale score moments within or across equatings. For example, 96_23 1-PL matches 94_23 1-PL under observed and true score equating

Table 18. Weight. Diff. Indices between Equipercentile and IRT Verbal Reasoning

		RMS	MAD	MSD	
96_23	Observed				
	EQU-1PL	**0.3625**	**0.1314**	-0.0430	1
	EQU-2PL	0.3579	0.1281	**-0.0506**	
	EQU-3PL	*0.2773*	*0.0759*	*-0.0218*	
	True				
	EQU-1PL	**0.4223**	**0.1784**	*0.0040*	2
	EQU-2PL	0.2789	0.0778	-0.0209	
	EQU-3PL	*0.1880*	*0.0353*	**0.0347**	
94_23	Observed				
	EQU-1PL	**0.3625**	**0.1314**	*-0.0430*	1
	EQU-2PL	0.3549	0.1259	-0.0485	
	EQU-3PL	*0.3531*	*0.1247*	**-0.0497**	
	True				
	EQU-1PL	**0.4223**	**0.1784**	*0.0040*	2
	EQU-2PL	0.3721	0.1385	**0.0398**	
	EQU-3PL	*0.2946*	*0.0868*	-0.0119	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 23. Number of -1 and +1 point differences between IRT rounded scale score equivalents and equipercentile equivalents - Verbal Reasoning

Section	Form	Method		-1	1
VR	96_23	Observed			
			1-PL	7	2
			2-PL	6	3
			3-PL	3	1
		True			
			1-PL	10	2
			2-PL	6	1
			3-PL	2	1
	94_23	Observed			
			1-PL	7	2
			2-PL	6	2
			3-PL	5	2
		True			
			1-PL	10	2
			2-PL	9	1
			3-PL	5	1

Table 24. Weighted Difference Indices for Verbal Reasoning Equivalents

				RMS	MAD	MSD	
24a.	IRT	96_23	Observed				
			1PL-2PL	*0.0872*	*0.0076*	*-0.0076*	
			1PL-3PL	**0.2334**	**0.0545**	0.0212	
			2PL-3PL	0.2262	0.0512	**0.0288**	
			True				
			1PL-2PL	0.3172	0.1006	*-0.0249*	
			1PL-3PL	**0.3791**	**0.1437**	0.0307	
			2PL-3PL	*0.3126*	*0.0977*	**0.0556**	
		94_23	Observed				
			1PL-2PL	0.0739	0.0055	-0.0055	
			1PL-3PL	**0.0821**	**0.0067**	**-0.0067**	
			2PL-3PL	*0.0356*	*0.0013*	*-0.0013*	
			True				
			1PL-2PL	*0.0399*	*0.0399*	0.0358	
			1PL-3PL	**0.3026**	**0.0916**	*-0.0158*	
			2PL-3PL	0.2273	0.0516	**-0.0516**	
24b.	Observe	96_23					
	- True		*1PL-1PL*	*0.2167*	*0.0470*	0.0470	1
			2PL-2PL	0.2474	0.0612	*0.0297*	
			3PL-3PL	**0.3112**	**0.0969**	**0.0565**	
		94_23					
			1PL-1PL	0.2167	0.0470	0.0470	1
			2PL-2PL	**0.2970**	**0.0882**	**0.0882**	
			3PL-3PL	*0.1946*	*0.0379*	*0.0379*	
24c.	96-	Observed					
			1PL-1PL	*0.0000*	*0.0000*	*0.0000*	2
			2PL-2PL	0.0462	0.0021	0.0021	
			3PL-3PL	**0.2185**	**0.0048**	**-0.0280**	
		True					
			1PL-1PL	*0.0000*	*0.0000*	*0.0000*	2
			2PL-2PL	0.2463	0.0607	**0.0607**	
			3PL-3PL	**0.3268**	**0.1068**	-0.0466	

Note: Bolded values indicate the highest indices within that equating, italicized values indicate the lowest indices within that equating, and superscript numbers indicate matched indices within or across equatings

Table 3. Observed Rounded Scale Score Equating Results for Biological Sciences Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	2	1	1	1
14	2	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	2
18	3	3	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	3
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	4	4	4	4
26	5	5	5	5
27	5	5	5	5
28	5	5	5	5
29	6	6	5	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	8	8	8	8
38	8	8	8	8
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	9	9	9	9
43	9	9	9	9
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	11
54	11	11	11	11
55	12	12	12	12
56	12	12	12	12
57	13	12	12	12
58	13	13	13	13
59	13	13	13	13
60	14	14	14	14
61	14	14	14	14
62	15	15	15	15
63	15	15	15	15

Table 4. True Rounded Scale Score Equating Results for Biological Sciences Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	2	1
14	2	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	2
18	3	3	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	3
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	4	4	4	4
26	5	5	5	5
27	5	5	5	5
28	5	5	5	5
29	6	6	5	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	8	8	8	8
38	8	8	8	8
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	9	9	9	9
43	9	9	9	9
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	11
54	11	11	11	11
55	12	12	12	12
56	12	12	12	12
57	13	12	12	12
58	13	13	13	13
59	13	13	13	13
60	14	14	14	14
61	14	14	14	14
62	15	15	15	15
63	15	15	15	15

Table 5. Observed Rounded Scale Score Equating Results for Biological Sciences Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	2	1	1	1
14	2	2	1	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	2
18	3	3	2	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	3
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	4	4	4	4
26	5	5	5	5
27	5	5	5	5
28	5	5	5	5
29	6	6	5	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	8	8	8	8
38	8	8	8	8
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	9	9	9	9
43	9	9	9	9
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	11
54	11	11	11	11
55	12	12	12	12
56	12	12	12	12
57	13	12	12	12
58	13	13	13	13
59	13	13	13	13
60	14	14	13	14
61	14	14	14	14
62	15	15	14	15
63	15	15	15	15

Table 6. True Rounded Scale Score Equating Results for Biological Sciences Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	2	2	1	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	2
18	3	3	2	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	3
22	4	4	3	4
23	4	4	4	4
24	4	4	4	4
25	4	4	4	4
26	5	5	5	5
27	5	5	5	5
28	5	5	5	5
29	6	6	5	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	8	8	8	8
38	8	8	8	8
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	9	9	9	9
43	9	9	9	9
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	10	11
52	11	11	11	11
53	11	11	11	11
54	11	11	11	11
55	12	12	12	12
56	12	12	12	12
57	13	12	12	12
58	13	13	13	13
59	13	13	13	13
60	14	14	13	14
61	14	14	14	14
62	15	15	14	15
63	15	15	15	15

Table 11. Observed Rounded Scale Score Equating Results for Physical Sciences Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	2
11	1	2	2	2
12	2	2	2	2
13	2	2	2	2
14	2	2	3	3
15	3	3	3	3
16	3	3	3	3
17	3	3	3	3
18	3	3	3	3
19	3	4	4	4
20	4	4	4	4
21	4	4	4	4
22	4	4	4	4
23	4	4	4	5
24	4	5	5	5
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	5	5	5	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	7	7	7	7
38	8	8	7	7
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	8	8	8	8
43	9	9	8	8
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	12
54	12	12	12	12
55	12	12	12	12
56	12	13	13	13
57	13	13	13	13
58	13	14	14	14
59	14	14	14	14
60	14	14	14	14
61	14	15	15	15
62	15	15	15	15
63	15	15	15	15

Table 12. True Rounded Scale Score Equating Results for Physical Sciences Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	2
11	1	1	2	2
12	2	2	2	2
13	2	2	2	2
14	2	2	3	3
15	2	3	3	3
16	3	3	3	3
17	3	3	3	3
18	3	3	3	3
19	3	3	4	4
20	4	4	4	4
21	4	4	4	4
22	4	4	4	4
23	4	4	4	5
24	4	5	5	5
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	5	5	5	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	7	7	7	7
38	8	8	7	7
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	8	8	8	8
43	9	9	8	8
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	12
54	12	12	12	12
55	12	12	12	12
56	12	13	13	13
57	13	13	13	13
58	13	14	14	14
59	14	14	14	14
60	14	14	14	14
61	15	15	15	15
62	15	15	15	15
63	15	15	15	15

Table 13. Observed Rounded Scale Score Equating Results for Physical Sciences Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	2
11	1	2	1	2
12	2	2	2	2
13	2	2	2	2
14	2	2	2	3
15	3	3	3	3
16	3	3	3	3
17	3	3	3	3
18	3	3	3	3
19	3	3	4	4
20	4	4	4	4
21	4	4	4	4
22	4	4	4	4
23	4	4	4	5
24	4	5	5	5
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	5	5	5	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	7	7	7	7
38	8	8	7	7
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	8	8	8	8
43	9	9	8	8
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	12
54	12	12	12	12
55	12	12	12	12
56	12	13	12	13
57	13	13	13	13
58	13	14	13	14
59	14	14	14	14
60	14	14	14	14
61	15	15	15	15
62	15	15	15	15
63	15	15	15	15

Table 14. True Rounded Scale Score Equating Results for Physical Sciences Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	2
11	1	1	1	2
12	2	2	2	2
13	2	2	2	2
14	2	2	2	3
15	2	3	2	3
16	3	3	3	3
17	3	3	3	3
18	3	3	3	3
19	3	3	3	4
20	4	4	4	4
21	4	4	4	4
22	4	4	4	4
23	4	4	4	5
24	4	5	5	5
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	5	5	5	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	6	6	6	6
33	6	6	6	6
34	7	7	7	7
35	7	7	7	7
36	7	7	7	7
37	7	7	7	7
38	8	8	7	7
39	8	8	8	8
40	8	8	8	8
41	8	8	8	8
42	8	8	8	8
43	9	9	8	8
44	9	9	9	9
45	9	9	9	9
46	9	9	9	9
47	10	10	10	10
48	10	10	10	10
49	10	10	10	10
50	10	10	10	10
51	11	11	11	11
52	11	11	11	11
53	11	11	11	12
54	12	12	12	12
55	12	12	12	12
56	12	13	13	13
57	13	13	13	13
58	13	14	13	14
59	14	14	14	14
60	14	14	14	14
61	15	15	15	15
62	15	15	15	15
63	15	15	15	15

Table 19. Observed Rounded Scale Score Equating Results for Verbal Reasoning Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	2	2
13	1	1	2	2
14	2	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	3
18	2	3	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	4	4
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	6	6	6	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	7	7	6	6
33	7	7	7	7
34	7	7	7	7
35	8	8	8	7
36	8	8	8	8
37	8	8	8	8
38	8	8	8	8
39	9	9	9	9
40	9	9	9	9
41	9	9	9	9
42	10	10	10	10
43	10	10	10	10
44	10	10	10	10
45	11	11	11	11
46	11	11	11	11
47	11	11	11	11
48	11	11	11	12
49	12	12	12	12
50	12	12	12	13
51	13	13	13	13
52	13	14	13	13
53	14	14	14	14
54	15	15	15	15
55	15	15	15	15

Table 20. True Rounded Scale Score Equating Results for Verbal Reasoning Form 23 to Form 15 1996

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	2	1
12	1	1	2	2
13	1	1	2	2
14	1	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	3	3
18	2	2	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	4	4
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	4	5	5	5
26	5	5	5	5
27	5	5	5	5
28	5	6	5	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	7	6	6	6
33	7	7	7	7
34	7	7	7	7
35	8	8	7	7
36	8	8	8	8
37	8	8	8	8
38	8	8	8	8
39	9	9	9	9
40	9	9	9	9
41	9	9	9	9
42	10	10	10	10
43	10	10	10	10
44	10	10	10	10
45	11	11	11	11
46	11	11	11	11
47	11	11	11	11
48	11	12	12	12
49	12	12	12	12
50	12	12	12	13
51	13	13	13	13
52	13	13	13	13
53	14	14	14	14
54	15	15	15	15
55	15	15	15	15

Table 21. Observed Rounded Scale Score Equating Results for Verbal Reasoning Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	2
13	1	1	2	2
14	2	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	3
18	2	3	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	4
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	5	5	5	5
26	5	5	5	5
27	5	5	5	5
28	6	6	6	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	7	7	7	6
33	7	7	7	7
34	7	7	7	7
35	8	8	8	7
36	8	8	8	8
37	8	8	8	8
38	8	8	8	8
39	9	9	9	9
40	9	9	9	9
41	9	9	9	9
42	10	10	10	10
43	10	10	10	10
44	10	10	10	10
45	11	11	11	11
46	11	11	11	11
47	11	11	11	11
48	11	11	11	12
49	12	12	12	12
50	12	12	12	13
51	13	13	13	13
52	13	13	13	13
53	14	14	14	14
54	15	15	15	15
55	15	15	15	15

Table 22. True Rounded Scale Score Equating Results for Verbal Reasoning Form 23 to Form 15 1994

Raw	Rounded			
Score	1-PL	2-PL	3-PL	Equi%ile
0	1	1	1	1
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	2
13	1	1	2	2
14	1	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	3
18	2	2	3	3
19	3	3	3	3
20	3	3	3	3
21	3	3	3	4
22	4	4	4	4
23	4	4	4	4
24	4	4	4	4
25	4	4	5	5
26	5	5	5	5
27	5	5	5	5
28	5	5	6	6
29	6	6	6	6
30	6	6	6	6
31	6	6	6	6
32	7	6	6	6
33	7	7	7	7
34	7	7	7	7
35	8	8	8	7
36	8	8	8	8
37	8	8	8	8
38	8	8	8	8
39	9	9	9	9
40	9	9	9	9
41	9	9	9	9
42	10	10	10	10
43	10	10	10	10
44	10	10	10	10
45	11	11	11	11
46	11	11	11	11
47	11	11	11	11
48	11	11	11	12
49	12	12	12	12
50	12	12	12	13
51	13	13	13	13
52	13	13	13	13
53	14	14	14	14
54	15	15	15	15
55	15	15	15	15

Figure 1. Observed Score Equating for Biological Sciences Form 23 to Form 15 1996

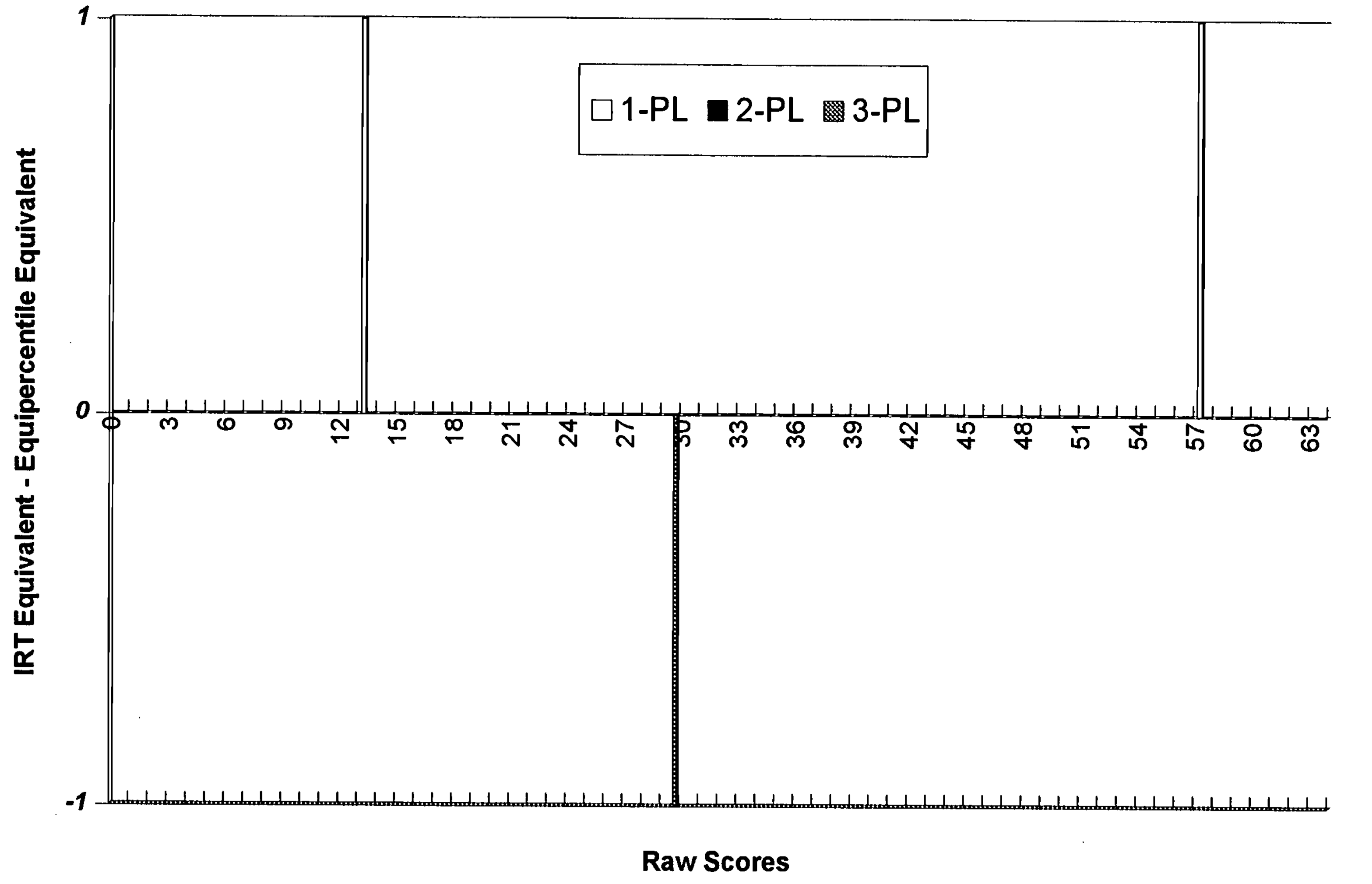


Figure 2. True Score Equating for Biological Sciences Form 23 to Form 15 1996

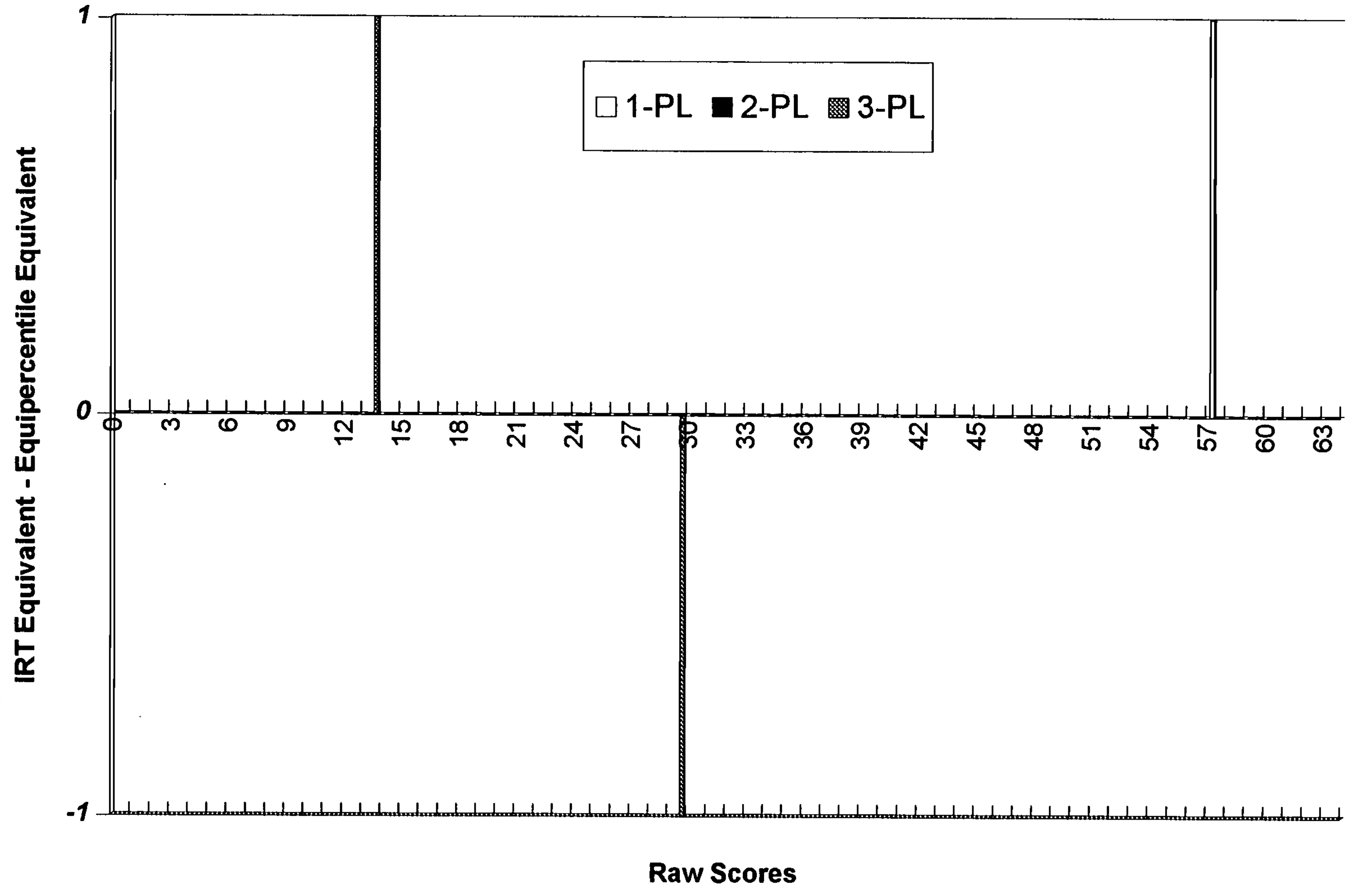


Figure 3. Observed Score Equating for Biological Sciences Form 23 to Form 15 1994

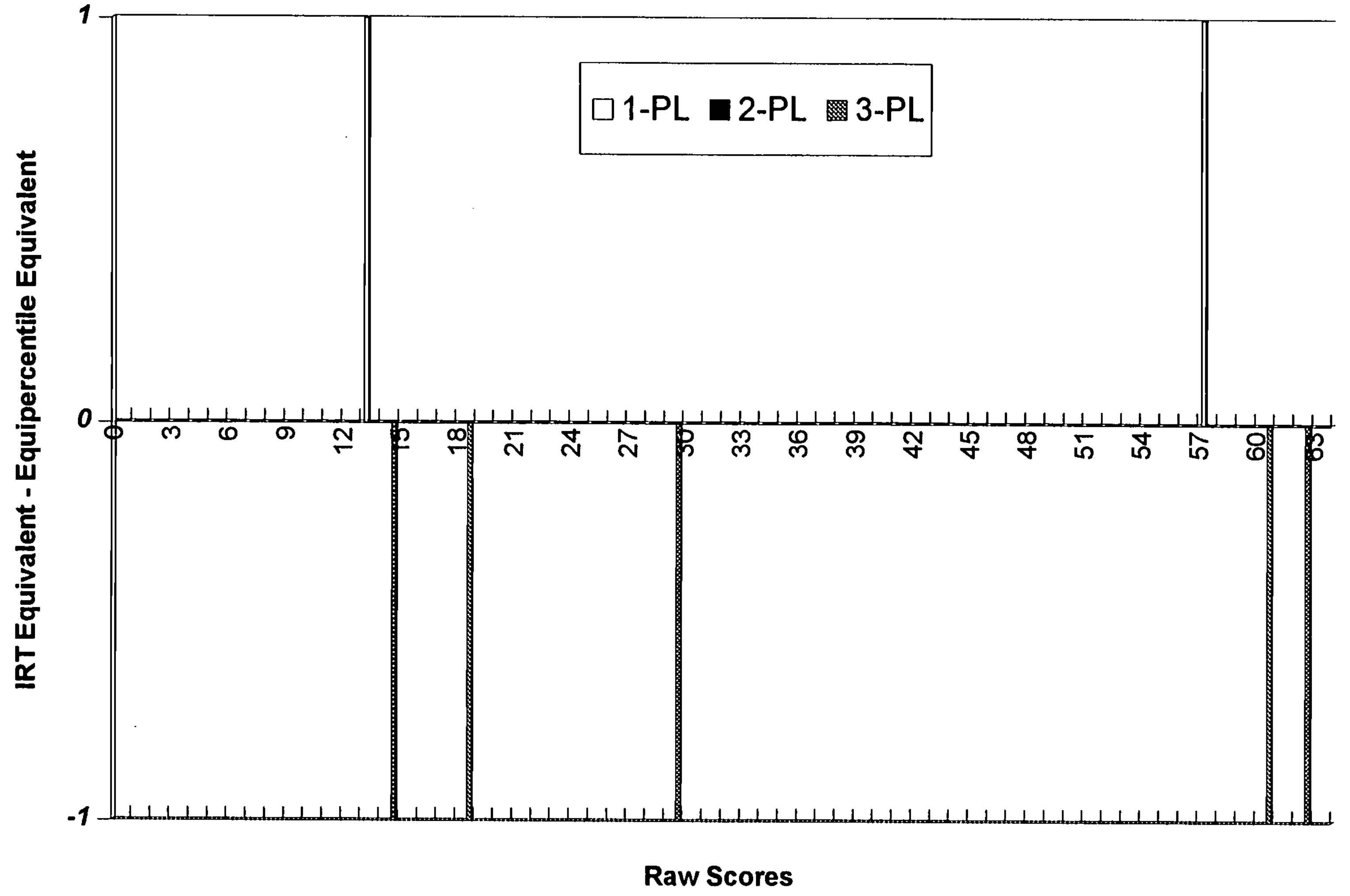


Figure 4. True Score Equating for Biological Sciences Form 23 to Form 15 1994

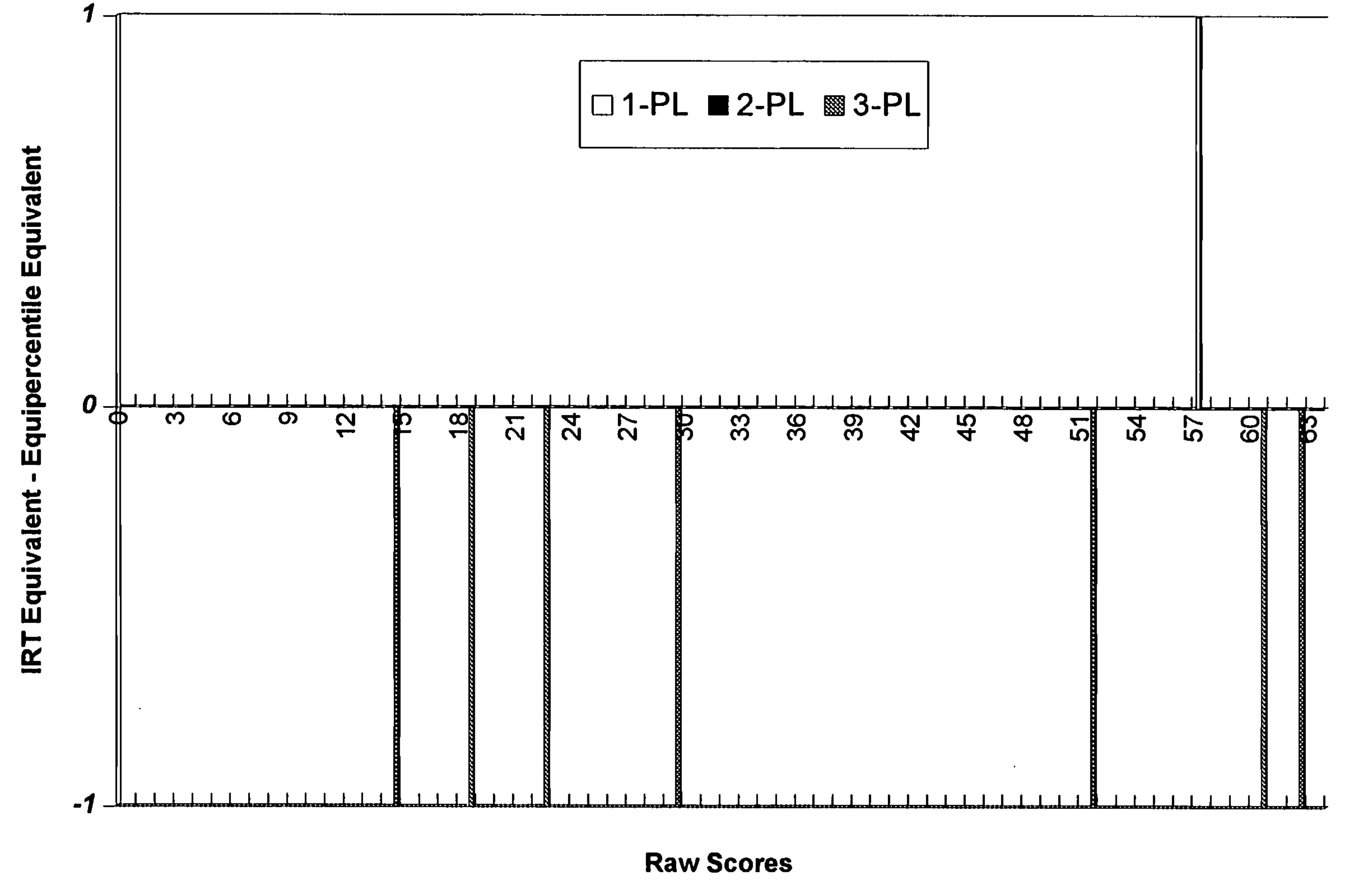


Figure 5. Observed Score Equating for Physical Sciences
Form 23 to Form 15 1996

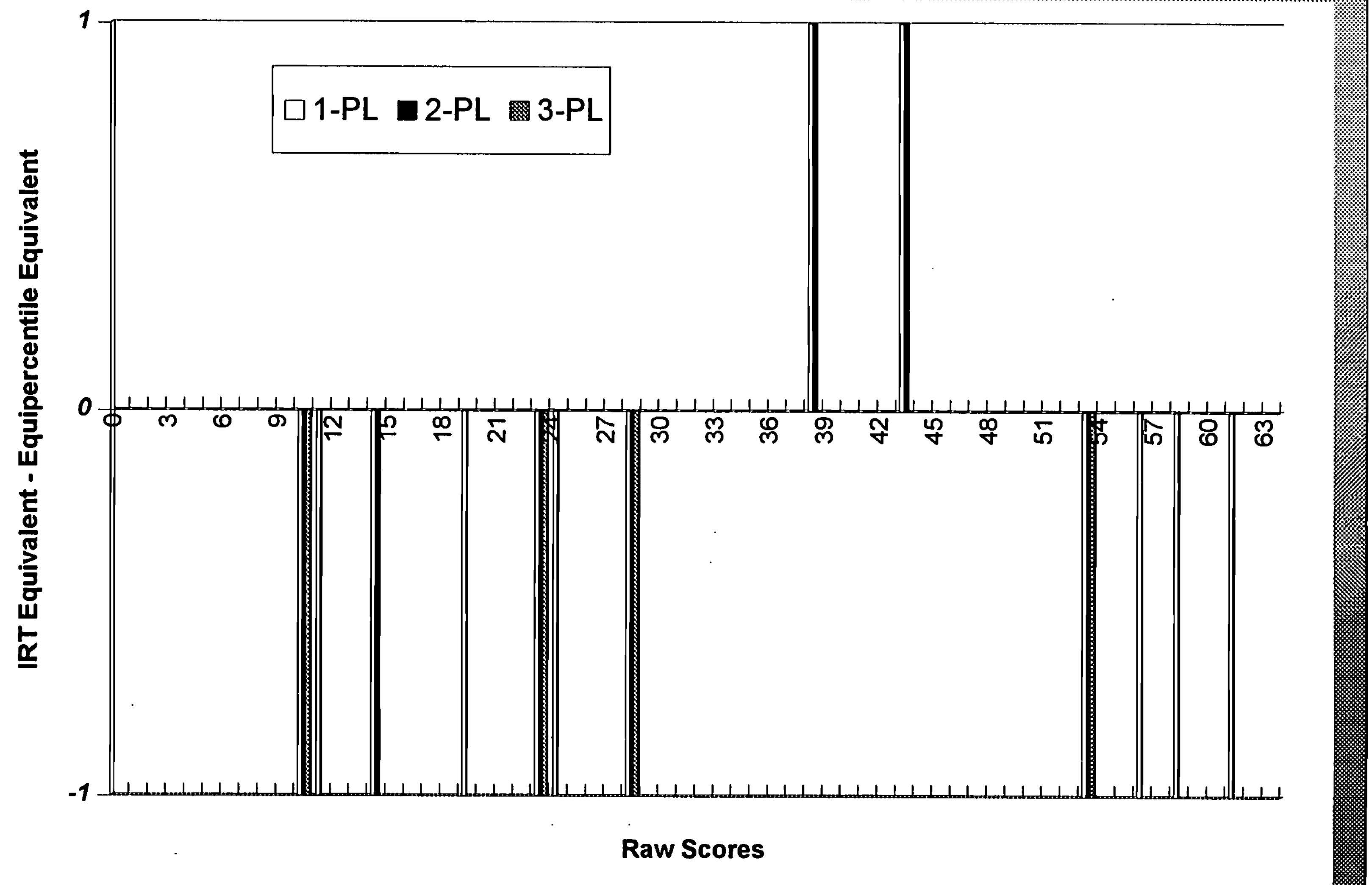


Figure 6. True Score Equating for Physical Sciences Form 23 to Form 15 1996

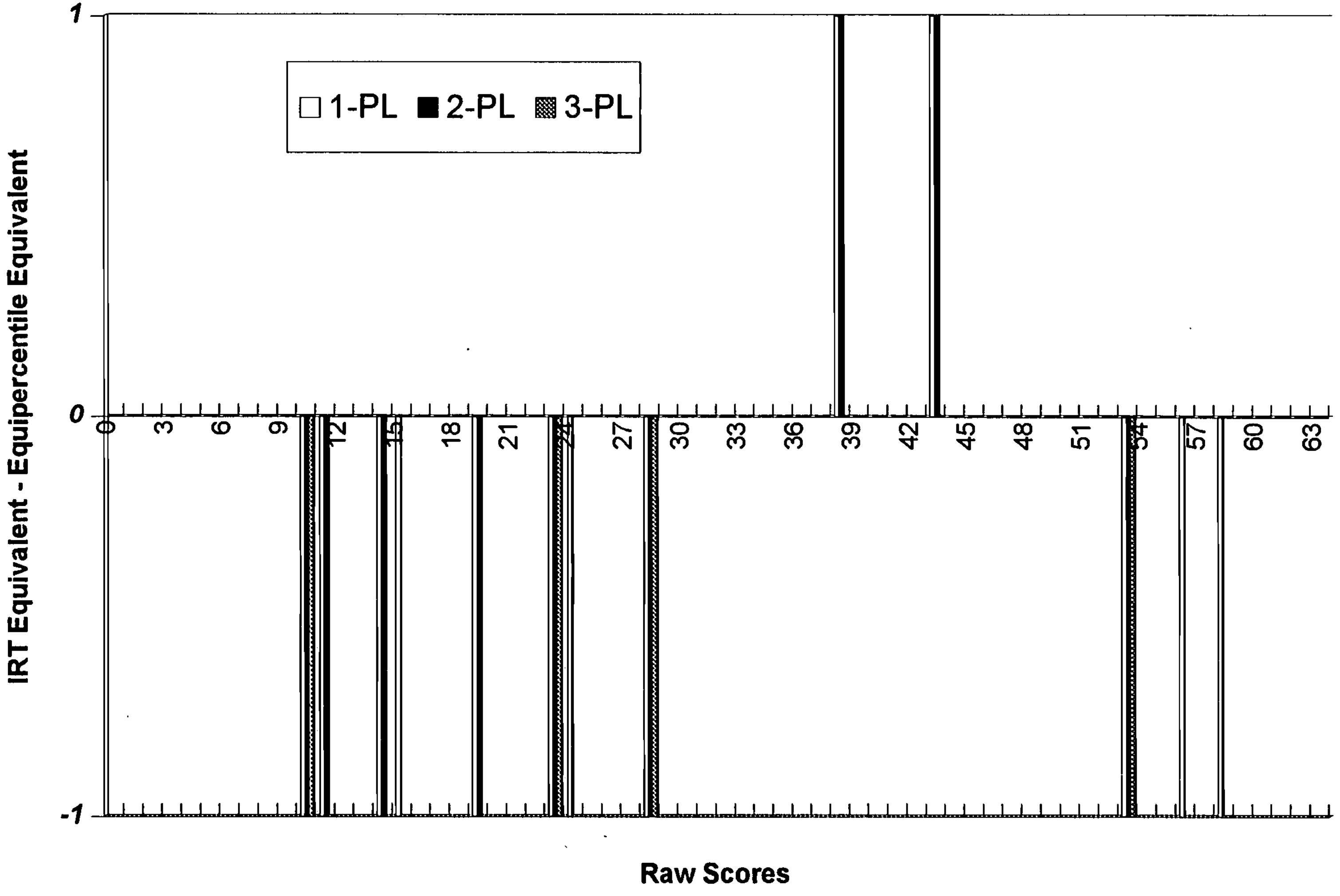


Figure 7. Observed Score Equating for Physical Sciences Form 23 to Form 15 1994

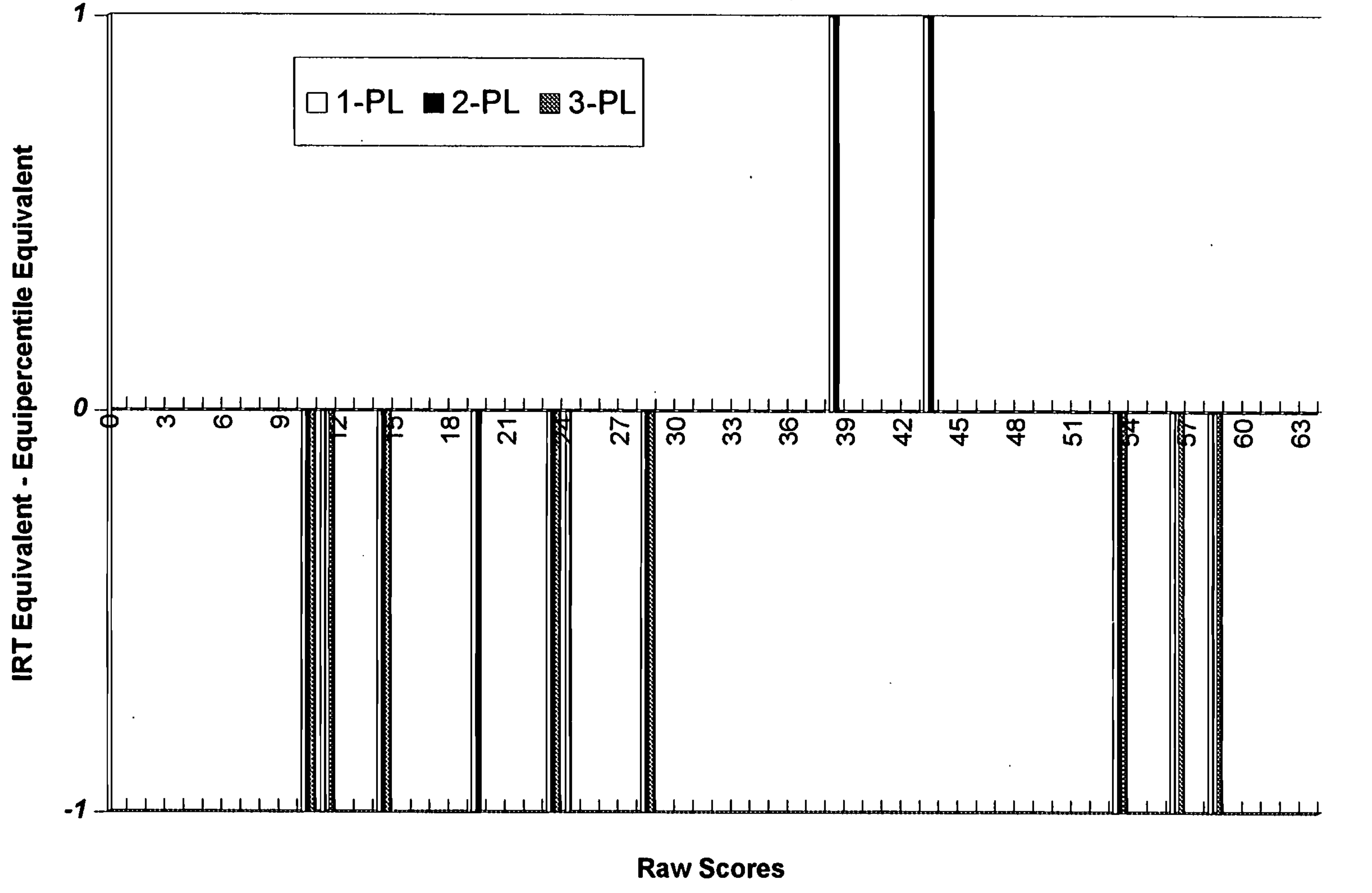


Figure 8. True Score Equating for Physical Sciences Form 23 to Form 15 1994

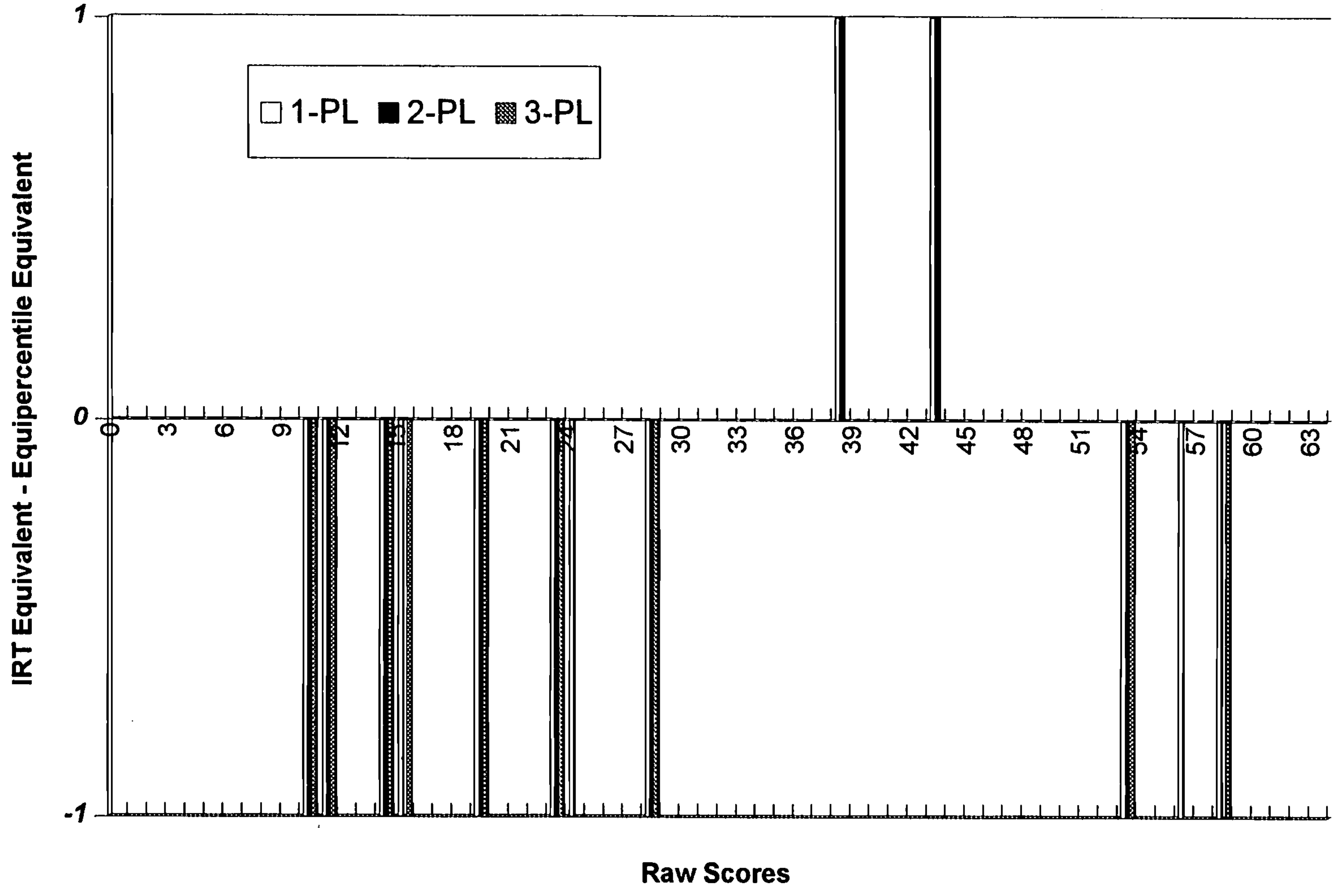


Figure 9. Observed Score Equating for Verbal Reasoning Form 23 to Form 15 1996

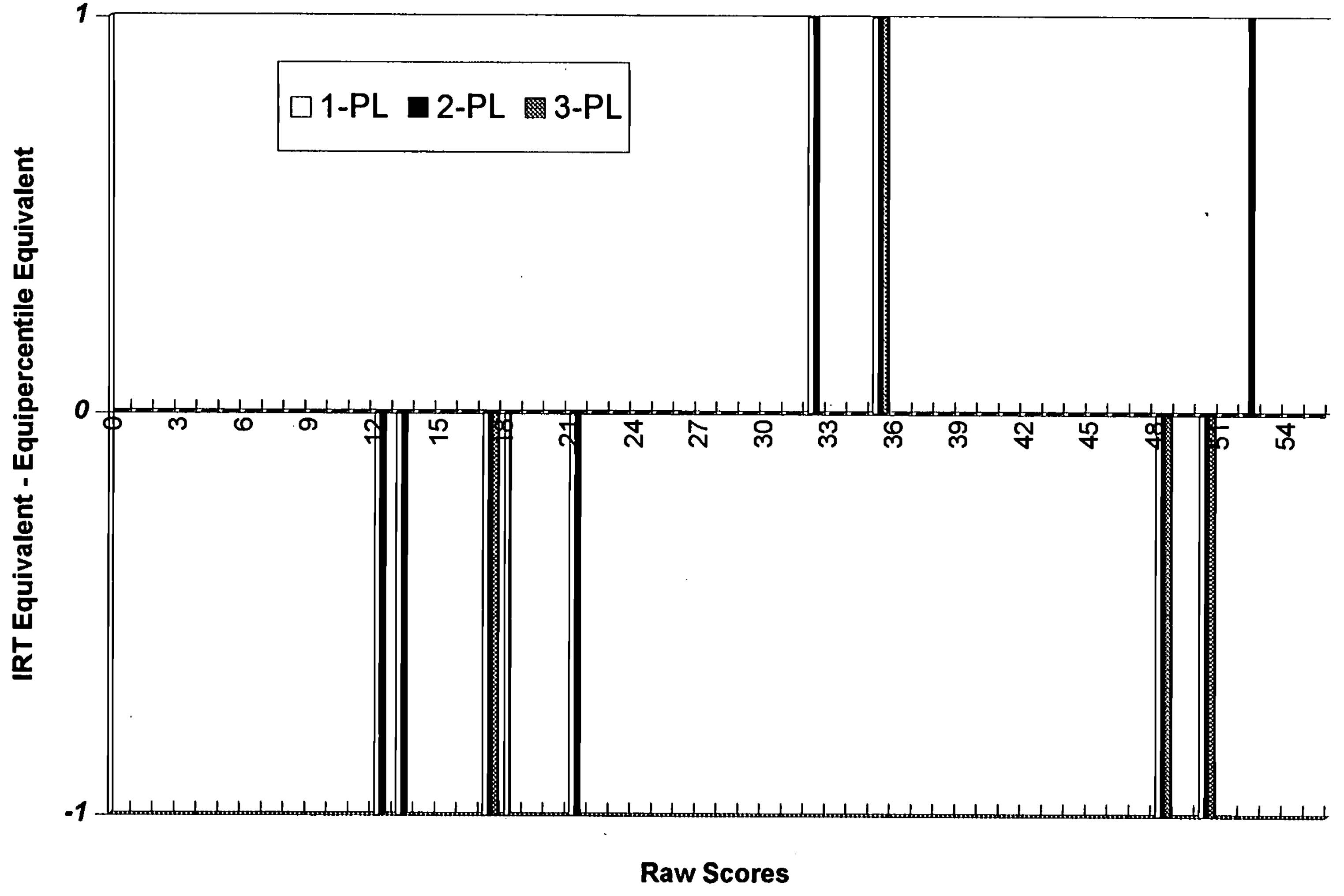


Figure 10. True Score Equating for Verbal Reasoning Form 23 to Form 15 1996

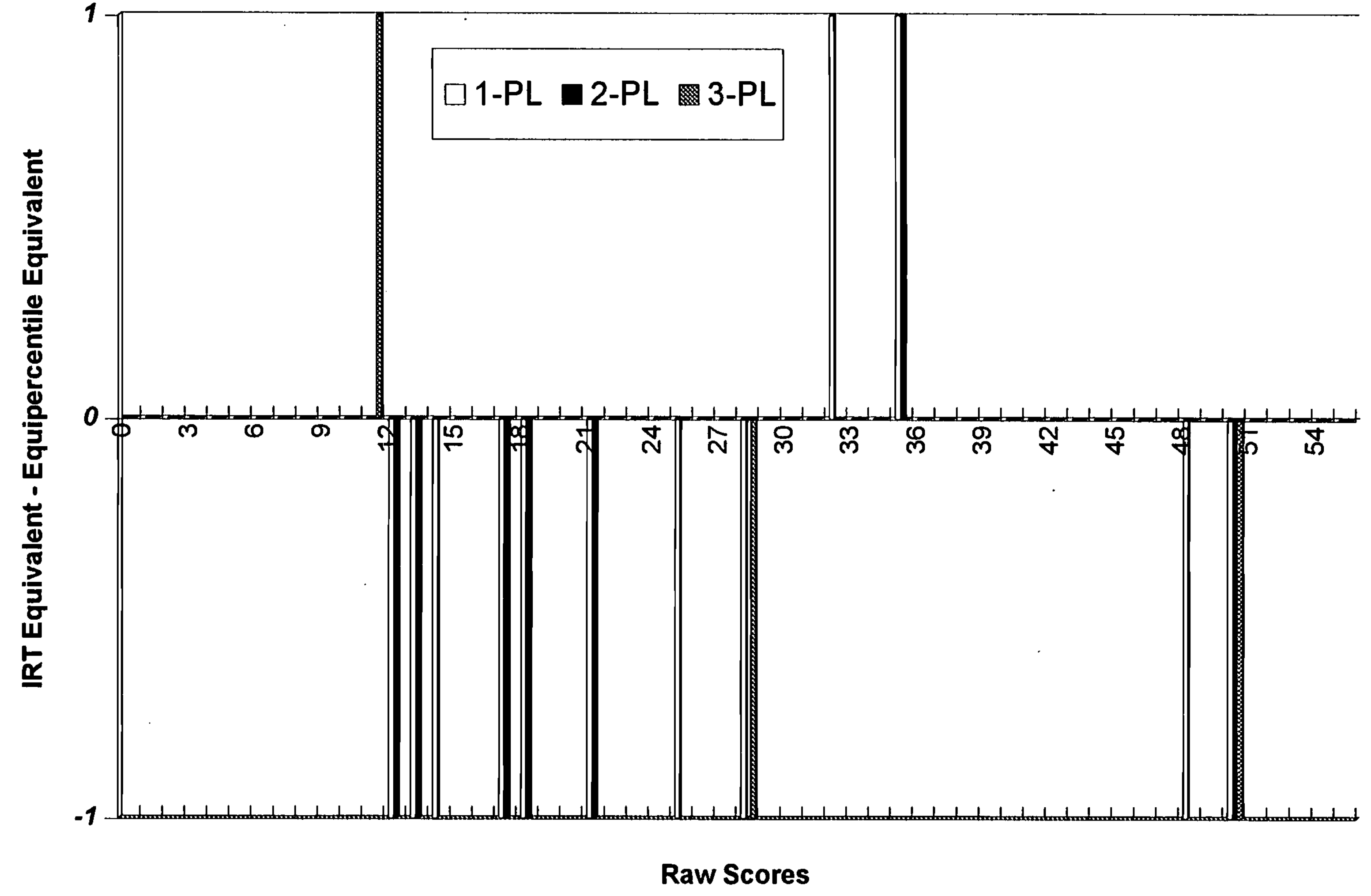


Figure 11. Observed Score Equating for Verbal Reasoning Form 23 to Form 15 1994

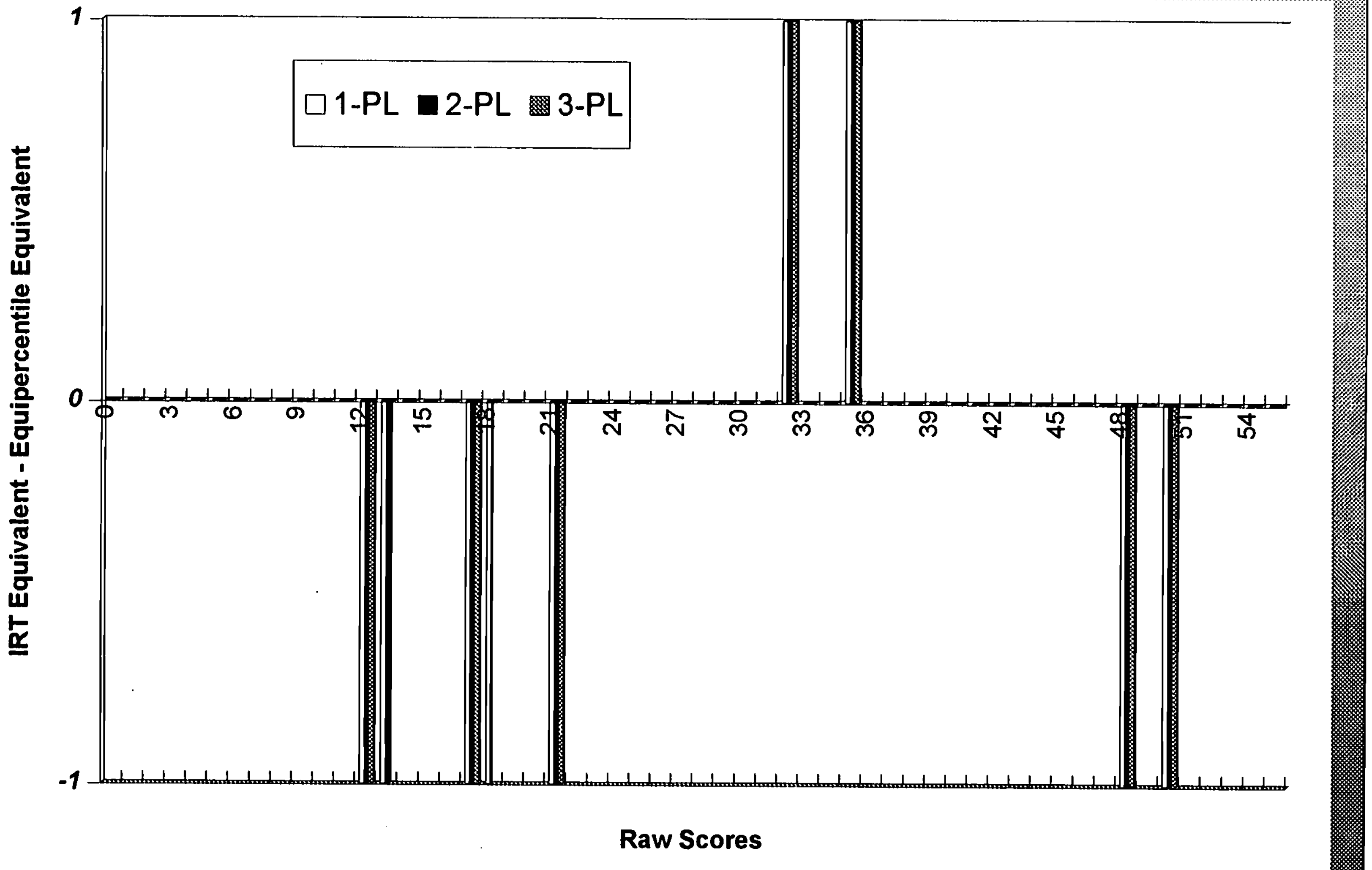
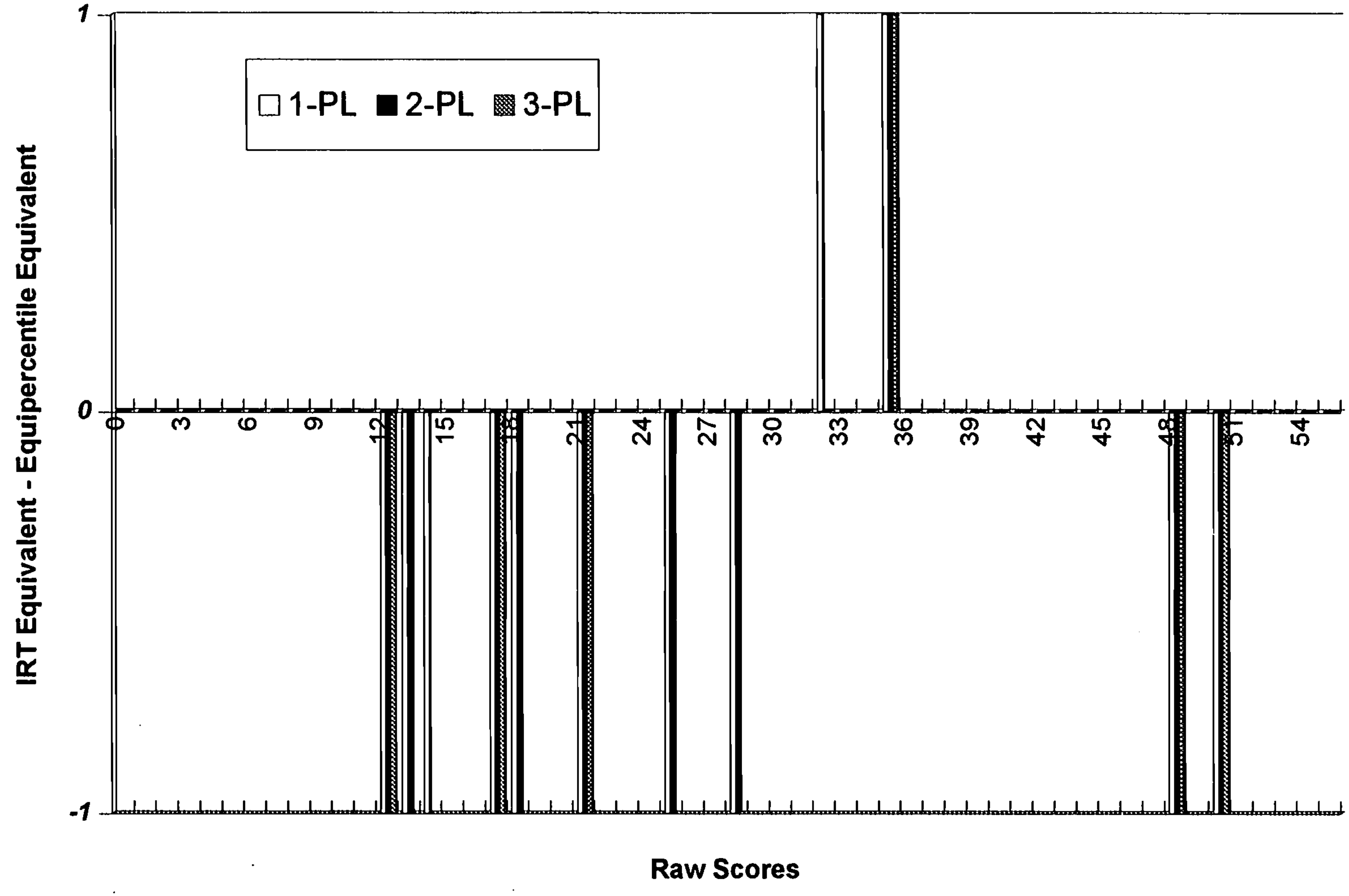


Figure 12. True Score Equating for Verbal Reasoning Form 23 to Form 15 1994



ASSOCIATION OF
AMERICAN
MEDICAL COLLEGES

2450 N STREET, NW, WASHINGTON, DC 20037
PHONE 202-828-0400 FAX 202-828-1125
HTTP://WWW.AAMC.ORG

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM033689

I. DOCUMENT IDENTIFICATION:

Title: IRT Equating of the MCAT

Author(s): Amy B. Hendrickson, Michael J. Kolen,

Corporate Source: The Association of American Medical Colleges, MCAT Division

Publication Date: Aug., 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
[✓ Level 1]	Level 2A	Level 2B
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:

Printed Name/Position/Title: Patricia Etienne, Ed.D. Director of MCAT Research

Organization/Address: 2450 N Street, NW
Washington, DC, 20037-1126

Telephone: (202) 828-0693

FAX: (202) 828-4799

E-Mail Address: petienne@aamc.org

Date: 2/8/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone:	**301-552-4200**
Toll Free:	**800-799-3742**
FAX:	**301-552-4700**
e-mail:	**ericfac@inet.ed.gov**
WWW:	**http://ericfacility.org**